Auditing Course Material

4. Production in Short Run

The assumptions of the law of diminishing returns are:

- **Fixed technology:** The law assumes that the production technology is fixed and does not change as the quantity of inputs changes.
- **Variable inputs:** The law assumes that at least one input in the production process is variable, while all other inputs are fixed.
- **Short-run analysis:** The law of diminishing returns is a short-run phenomenon. It assumes that the fixed inputs, such as capital and technology, cannot be varied in the short run.
- **Homogeneous inputs:** The law assumes that all units of the variable input are identical in terms of their productivity and quality.
- **Constant scale of production:** The scale of production is assumed to be constant. This means that the firm cannot expand its plant size or add more machinery to its production process.
- **Law of diminishing marginal utility:** The law of diminishing marginal utility is a related assumption of the law of diminishing returns. It assumes that as more and more units of a good are consumed, the marginal utility (or satisfaction) from each additional unit consumed will eventually decline.

---

4. Production in Short Run

The law of diminishing marginal returns states that as the quantity of a variable input, such as labour or raw materials, increases while other inputs remain constant, the additional output eventually diminishes. However, there is an exception to this rule, which is called **network effects**. For example, in the case of Microsoft Office and Outlook, the more people who use the software, the more valuable it becomes for new users. This is because the software's installed base creates a larger network of compatible users, making it more useful for collaboration, communication, and data sharing. As the number of users increases, the value of the software also increases, which means Microsoft's marketing efforts become more productive, leading to even more growth in the user base. This positive feedback loop is called a network effect, and it is an exception to the law of diminishing marginal returns.

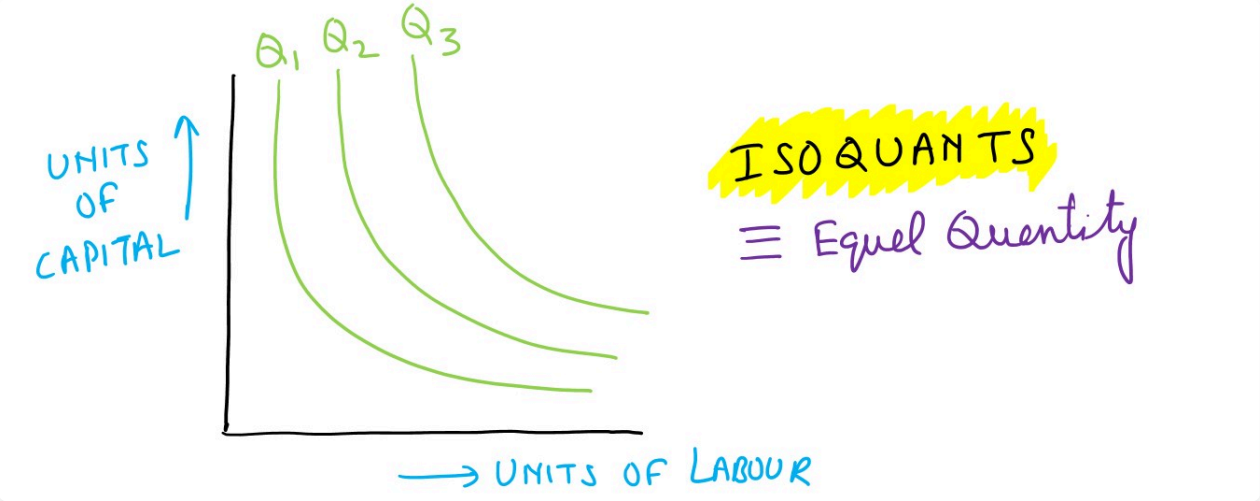
---

5. Production in Long Run

Let us now understand, how the firm should be combining resources for optimal output in long run, where there are no fixed resources and all resources can be varied.

---

5. Production in Long Run



Isoquants are curves that represent all the possible combinations (**technologically efficient**) of two inputs that can produce a fixed level of output. In other words, isoquants show the different combinations of labour and capital that can produce the same level of output. The term "isoquant" comes from the Greek word "iso" meaning equal and "quant" meaning quantity, so the term "isoquant" means equal quantity curves.

The production function describes the maximum output that can be produced from a given set of inputs, while the isoquant shows all the different combinations of inputs that can produce a given level of output.

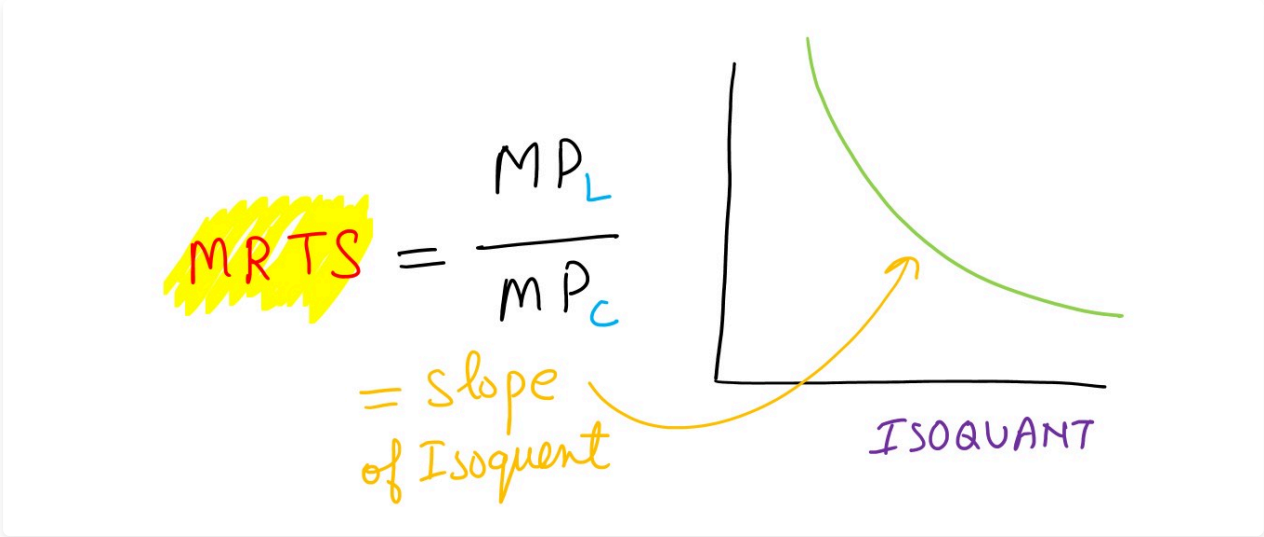
Along a particular isoquant, such as Q_1, the rate of output produced remains constant, but the combination of resources varies. Isoquants farther from the origin Q_3 represent greater output rates than Isoquants nearer to the origin Q_2.

Isoquants have the following properties:

- **Isoquants never intersect:** Two isoquants cannot cross because that would mean that the same level of output can be produced by two different combinations of inputs, which violates the assumption of fixed output.
- **Isoquants slope downwards and thus have a negative slope:** The slope of the isoquant represents the rate at which one input can be substituted for the other while maintaining the same level of output. As we move along the isoquant from left to right, we are replacing one input with the other. The slope of the isoquant becomes steeper as we move to the right, indicating that more of one input is needed to offset the loss of the other input.
- **Isoquants are convex to the origin:** The curvature of the isoquant represents the rate at which one input can be substituted for the other. As we move along the isoquant, the slope becomes steeper, indicating that more of one input is needed to offset the loss of the other input. The curvature of the isoquant is related to the concept of diminishing marginal returns.

---

5. Production in Long Run



The slope of an isoquant measures the ability of additional units of one resource—in this case, labour—to substitute in production for another—in this case, capital. As noted already, the isoquant has a negative slope. The absolute value of the slope of the isoquant is the marginal rate of technical substitution, or MRTS, between two resources. The MRTS is the rate at which labour substitutes for capital without affecting output.

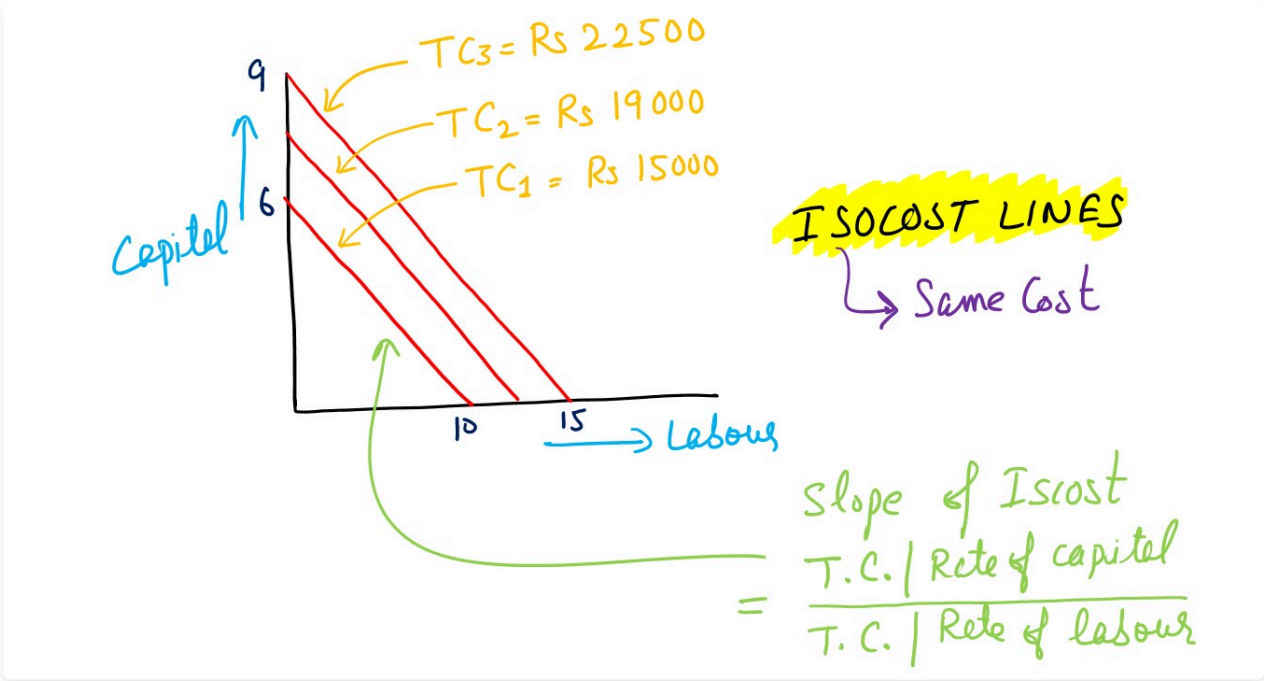
The extent to which one input substitutes for another, as measured by the marginal rate of technical substitution, is directly linked to the marginal productivity of each input. Anywhere along the isoquant, the marginal rate of technical substitution of labour for capital equals the marginal product of labour divided by the marginal product of capital, which also equals the absolute value of the slope of the isoquant.

MRTS = Magnitude of slope of isoquant = $\frac{Marginal \ productivity \ of \ labour}{Marginal \ productivity \ of \ capital} = \frac{MP_L}{MP_C}$

---

5. Production in Long Run

Isoquants graphically illustrate a firm's production function for all quantities of output the firm could possibly produce. We turn now to the question of what combination of resources to employ to minimize the cost of producing a given rate of output. It depends on the cost of resources.



Suppose a unit of labour costs (w) the firm Rs 1,500 per month, and a unit of capital (r) costs Rs 2,500 per month. The total cost (TC) of production per month is:

TC = (w × L) + (r × C) = 1,500 L + 2,500 C

where w is the monthly wage rate, L is the quantity of labour employed, r is the monthly cost of capital, and C is the quantity of capital employed.

An **isocost line** identifies all combinations of capital and labour the firm can hire for a given total cost. Again, iso is Greek for "equal," so an isocost line is a line representing resource combinations of equal cost.

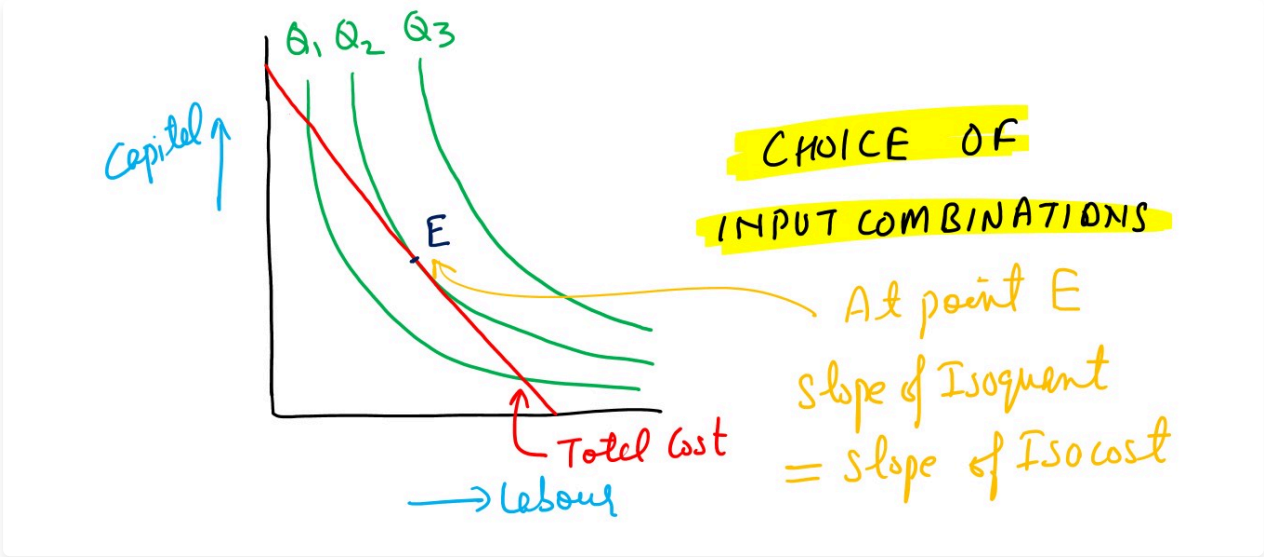
In the figure, for example, the line TC_1= Rs 15,000 identifies all combinations of labour and capital that cost the firm Rs 15,000 per month. The entire Rs 15,000 could pay for either 6 units of capital or 10 units of labour per month. Or the firm could employ any other combination of resources along the isocost line. Higher costs are represented by isocost lines farther from the origin.

The slope of each equals the negative of the monthly wage (w) rate divided by the rental cost of capital per month (r).

Slope of Isocost = $\- \frac{Total \ Cost/ rate \ of \ capital}{Total \ cost/ rate \ of \ labour}= - \frac{w}{r}$

5. Production in Long Run

Let us now bring together the isoquants and the isocost lines in a figure. For a required output (say Q_2), because the profit-maximizing firm wants to produce its chosen output at the minimum cost, it tries to find the isocost line closest to the origin that still touches the isoquant. The isoquant for required units of output is tangent to the isocost line at point E. From that point of tangency, any movement in either direction along an isoquant increases the cost. So the tangency between the isocost line and the isoquant shows the minimum cost required to produce a given output.



At point E in the figure, the isoquant and the isocost line have the same slope. As mentioned already, the absolute value of the slope of an isoquant equals the marginal rate of technical substitution (MRTS) between labour and capital, and the absolute value of the slope of the isocost line equals the ratio of the input prices. So when a firm produces output in the least costly way, the marginal rate of technical substitution must equal the ratio of the resource prices.

MRTS = \(\frac{MP_L}{MP_C} = \frac{w}{r} \)

which can also be re-written as:

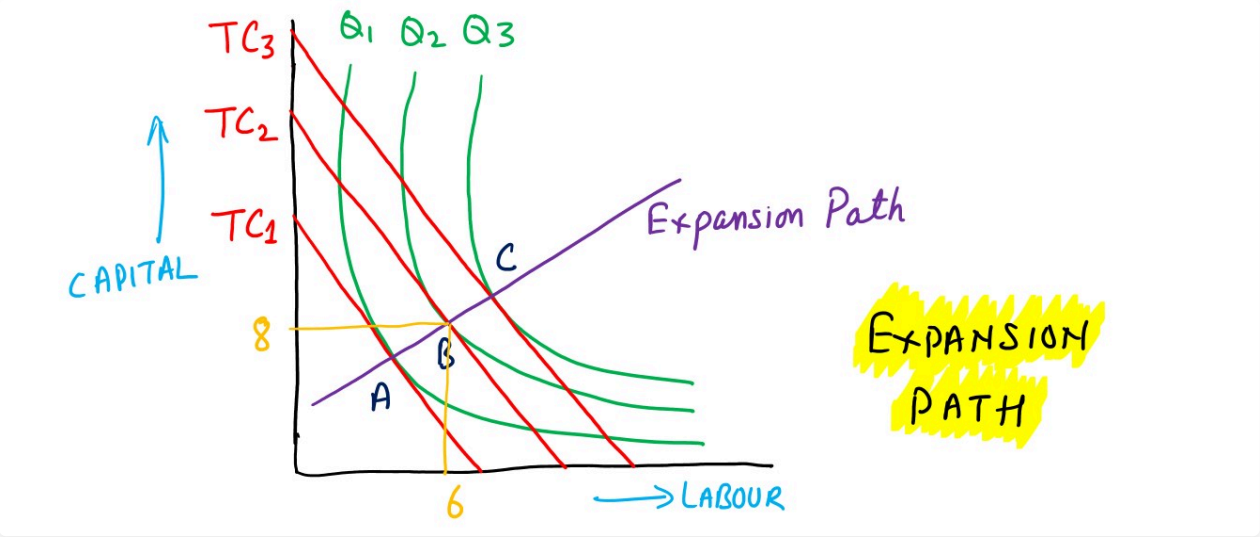
\(\frac{MP_L}{w} = \frac{MP_C}{r} \)

This equality shows that the firm adjusts resource use so that the rate at which one input substitutes for another in production—that is, the marginal rate of technical substitution—equals the rate at which one resource exchanges for another in resource markets, which is w/r. If this equality does not hold, the firm could adjust its input mix to produce the same output for a lower cost.

---

5. Production in Long Run

Imagine a set of isoquants representing each possible rate of output. Given the relative cost of resources, we could then draw isocost lines to determine the optimal combination of resources for producing each rate of output. The points of tangency in the figure show the least-cost input combinations for producing several output rates. For example, output rate Q_2 can be produced most cheaply using 8 units of capital and 6 units of labour.



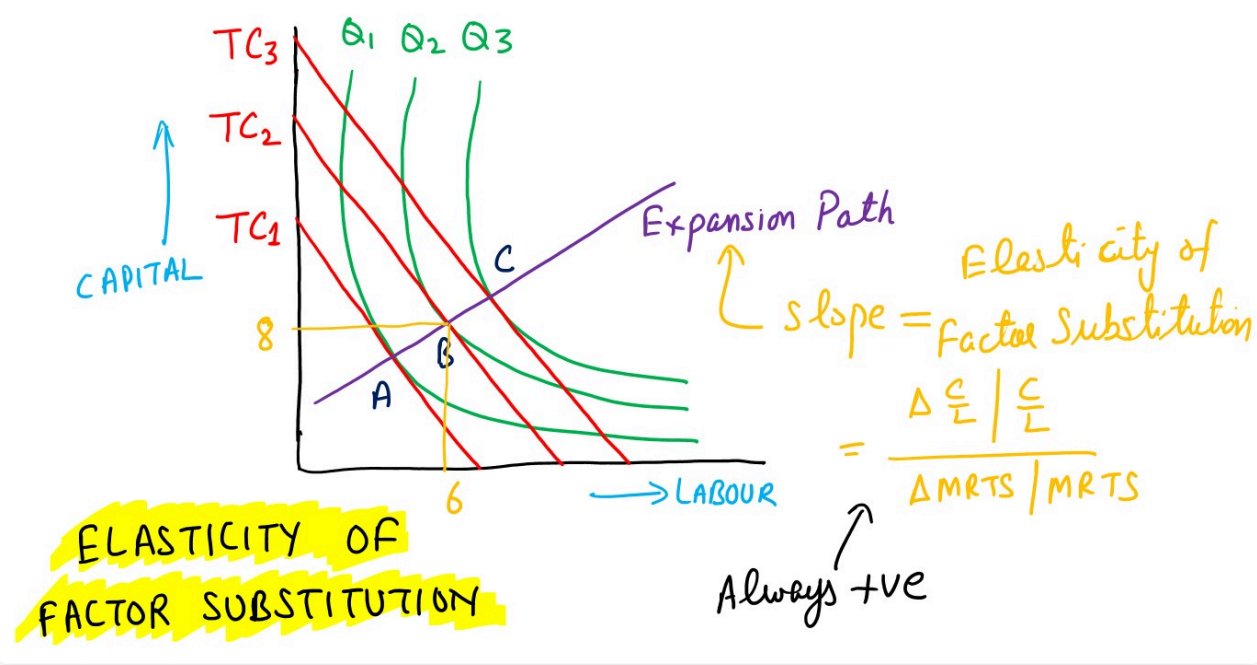
The line formed by connecting these tangency points is the firm's **expansion path**. The expansion path need not be a straight line, although it will generally slope upward, indicating that the firm will expand the use of both resources in the long run as output increases.

Using the expansion path, the company can make long-term planning decisions related to input usage. For example, if the company expects to increase its production level in the future, it can use the expansion path to determine the optimal input combination needed to produce the higher level of output at the lowest possible cost. The company can also use the expansion path to analyze the impact of changes in input prices on its production cost.

The slope of the expansion path indicates the **elasticity of factor substitution** between the two inputs. A steep slope indicates that the company can easily substitute one input for the other, while a flatter slope indicates that substitution is difficult.

---

5. Production in Long Run



Elasticity of factor substitution is a measure of the degree to which a company can substitute one input for another in the production process. While the marginal rate of technical substitution (MRTS) indicates the rate at which one input can be substituted for another along an isoquant, it does not reveal how difficult or easy it is to substitute one input for another.

The elasticity of substitution is a better measure of the substitutability of factors. It is defined as the percentage change in the capital-labour ratio (C/L) divided by the percentage change in marginal rate of technical substitution (MRTS), i.e.

$$ σ = \frac{\% \ change \ in \ C/L}{\% \ change \ in \ MRTS} $$

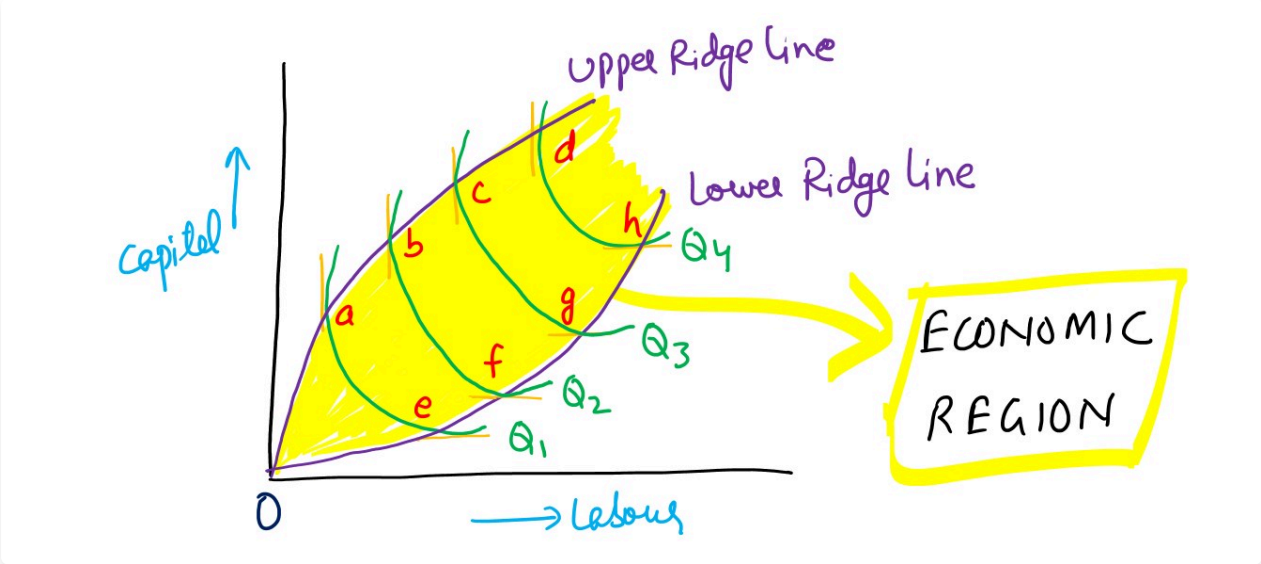
The value of elasticity of substitution is always positive (between zero to infinity), and it is independent of the units of measurement of capital and labour.

For example, consider a shoe manufacturer that uses both labour and capital in its production process. The elasticity of substitution would measure how easily the company can substitute one input for the other. If the elasticity of substitution is high, the company can easily substitute labour for capital or vice versa, and this can impact the company's production cost.

---

5. Production in Long Run

It is noteworthy that the whole isoquant map or production plane is not technically efficient, nor is every point on isoquant technically efficient. The reason is that, on a convex isoquant, the MRTS decreases along the isoquant. The limit to which MRTS can decrease, is zero.

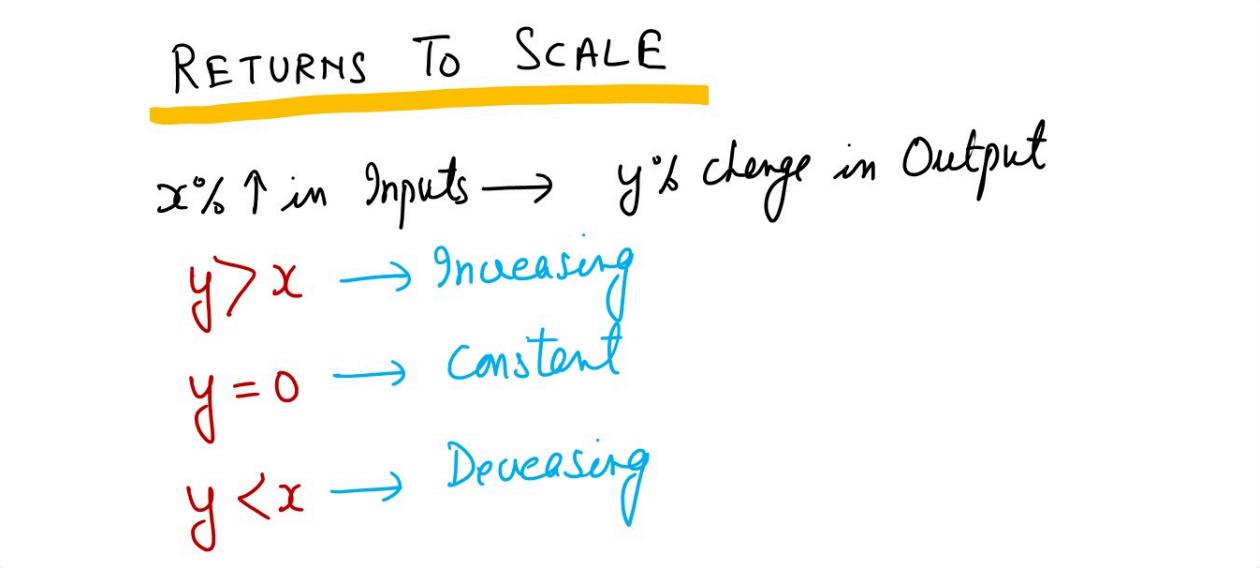


A zero MRTS implies that there is a limit to which one input can substitute another. It also determines the minimum quantity of an input which must be used to produce a given output. Beyond this point, an additional employment of one input will necessitate employing additional units of the other input. Such a point on an isoquant may be obtained by drawing a tangent to the isoquant and parallel to the vertical and horizontal axes, as shown by dashed lines in the figure. By joining the resulting points a, b, c and d, we get a line called the **upper ridge line**, Od. Similarly, by joining the points e, f, g and h, we get the **lower ridge line**, Oh. The ridge lines are locus of points on the isoquants where the marginal products (MP) of the inputs are equal to zero. The upper ridge line implies that MP of capital is zero along the line, Od. The lower ridge line implies that MP of labour is zero along the line, Oh.

The area between the two ridge lines, Od and Oh, is called **Economic Region or technically efficient region** of production. Any production technique, i.e., capital-labour combination, within the economic region is technically efficient to produce a given output. Any production technique outside this region is technically inefficient since it requires more of both inputs to produce the same quantity.

---

6. Returns to Scale



The concept of returns to scale explains the behaviour of output in response to a proportional and simultaneous change in ALL inputs. If all inputs are changed at the same time, and suppose are increased proportionately, then the concept of returns to scale has to be used to understand the behaviour of output.

The laws of returns to scale always refer to the long run because only in the long run are all the factors of production variable. In other words, only in the long run is it possible to change all the factors of production.

Returns to scale are classified as follows:

1. **Increasing Returns to Scale (IRS):** If output increases more than proportionate to the increase in all inputs.
2. **Constant Returns to Scale (CRS):** If all inputs are increased by some proportion, output will also increase by the same proportion.
3. **Decreasing Returns to Scale (DRS):** If increase in output is less than proportionate to the increase in all inputs.

For example, if all factors of production are doubled and output increases by more than two times, then the situation is of **increasing** returns to scale. On the other hand, if output does not double even after a cent per cent increase in input factors, we have **diminishing** returns to scale.

The laws of returns to scale may be explained more precisely through a production function. Let us assume a production function involving two variable inputs (K and L) and one commodity X. The production function may then be expressed as

$Q = f(K, L)$

If land, K, and labour, L, are both multiplied by h and Q increases by k, we get,

$kQ = f(hK, hL)$

We have constant, increasing or decreasing returns to scale, respectively depending upon, whether $k = h$, $k > h$ or $k < h$.

Accordingly, it reveals 3 laws of returns to scale:

1. If $h = k$, production function reveals constant returns to scale.
2. If $h > k$, it reveals decreasing returns to scale.
3. If $h < k$, it reveals increasing returns to scale.

Increasing returns to scale arise because as the scale of operation increases, a greater division of labour and specialization can take place and more specialised and productive machinery can be used. Decreasing returns to scale arise primarily because as the scale of operation increases, it becomes more difficult to manage the larger firm. The primary cause behind Increasing Returns to Scale is called **Economics of Scale** and cause behind Diminishing returns to Scale is called **Diseconomics of Scale**.

---

7. Important Production Functions

Cobb-Douglas Production Function

In 1928, **Cobb** and **Douglas** introduced a famous two-factor production function, called **Cobb-Douglas production function**, in order to describe the distribution of the national income by help of production functions.

$Q = AK^{\alpha}L^{\beta}$

where:
Q = Output
A = Efficiency
K = Capital
L = Labour parameters
α and β represent the elasticity co-efficient of output for inputs K and L

Return to scale for Cobb-Douglas production function:

1. Constant Return to Scale→ α+ β =1
2. Increasing Return to Scale→ α + β >1
3. Decreasing Return to Scale→ α + β <1

Other Production Functions

In addition to the Cobb-Douglas production function, there are several other forms of production function, viz., Constant Elasticity Substitution (CES), Variable Elasticity of Substitution (VES), Leontief-type, and linear-type.

A **linear production function** is of the following form:

Q = aL + bK

Where Q is total Product, a is the productivity of L units of labour, b is the productivity of K units of capital.

The **CES production function** (was given in 1961 by *Arrow et al.*) is expressed as

$$Q = A[\alpha K^{- \beta} +(1- \alpha)L^{- \beta}]^ \frac{-1}{ \beta}$$

An important property of the CES production function is that it is homogeneous of degree 1. If β is assumed to be a variable, then the above function may be called the variable elasticity of substitution, VES function.

Leontief production function (fixed proportion production function) is useful when labour and capital must be furnished in a fixed proportion. The equation for a fixed proportion function is as follows:

Q = min (aK, bL)

where Q is the total Product, a and b are the coefficient of production of capital and labour respectively and K and L represent the units of capital and labour respectively. Another way to write Leontief Production function is

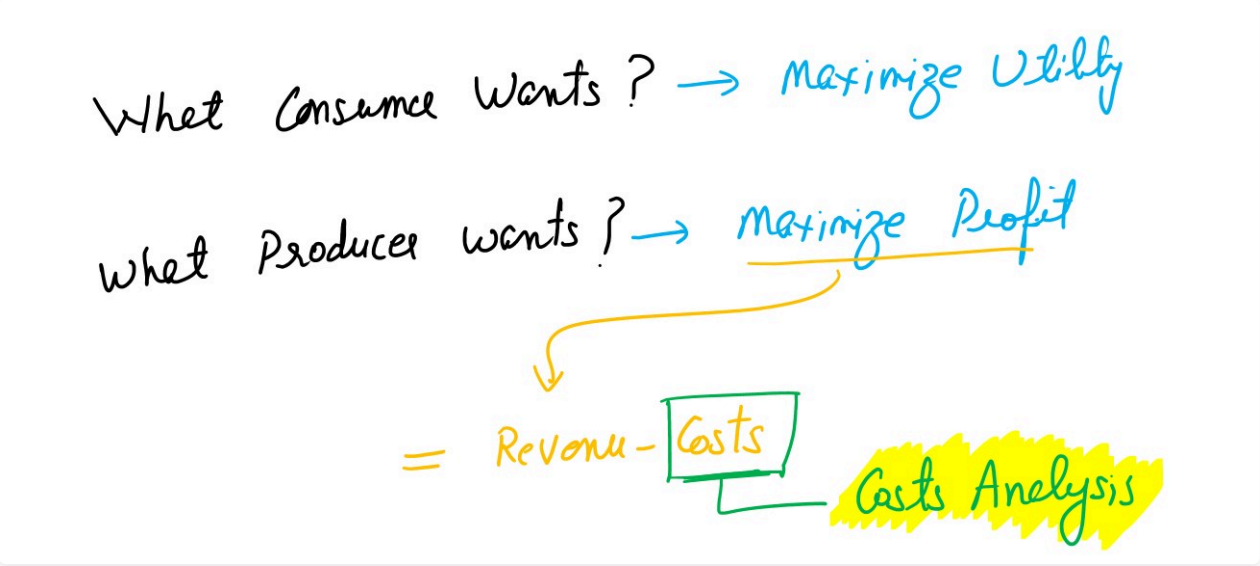
$$Q = min (\frac{L}{ \alpha}, \frac{K}{ \beta})$$

If $\frac{L}{ \alpha} < \frac{K}{ \beta}$, then $Q = \frac{L}{ \alpha}$, and L is considered as the binding constraint in the production process. This is because an increase in labour input is necessary to increase output, but an increase in K in such a situation will not increase output.

Conversely, if $\frac{L}{ \alpha} > \frac{K}{ \beta}$, then $Q = \frac{K}{ \beta}$ and K becomes the limitative factor or binding constraint on output. So, any additional increase in labour will not lead to an increase in output. If $\frac{L}{ \alpha} = \frac{K}{ \beta}$, both the factors are fully utilized. Thus, the total product under the fixed proportions production function is restricted by the lower of labour and capital.

---

1. Introduction



When evaluating the performance of a business, it is crucial to consider not only its revenues but also the costs incurred in generating those revenues. Managers conduct cost analysis to control expenses, plan strategically, and optimize resource allocation.

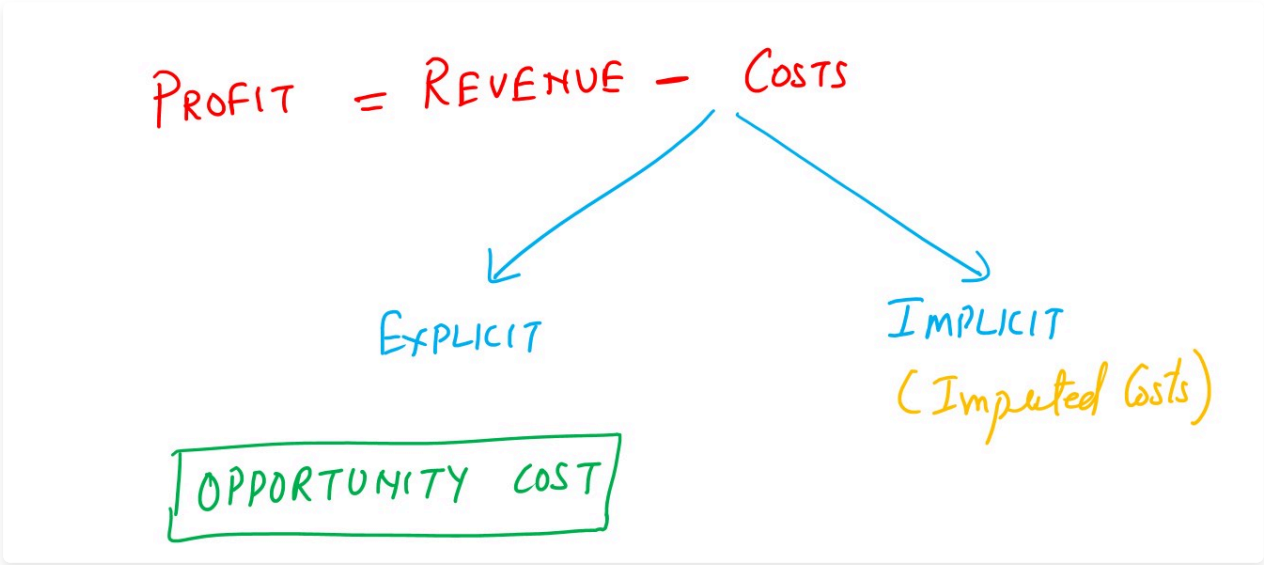
Let us start with the understanding of various types of costs.

---

1. Introduction

The cost of hiring a resource is determined by their opportunity cost, which is the value of the resource in its best alternative use. The **Opportunity cost** refers to the value of the next best alternative that is forgone when a decision is made. It is the cost of the best alternative that must be given up to pursue a certain action or decision.

Some firms own their resources and do not make direct cash payments. For example, a company may operate in a building it owns without paying rent, and small business owners may not pay themselves an hourly wage. Although no cash payment is made, these resources still have an opportunity cost since they could be used in other ways.



A firm's **explicit costs** are its cash payments for resources, such as wages, rent, interest, insurance, taxes, and more. On the other hand, **implicit costs** are the opportunity costs of utilizing resources owned or provided by the firm's owners, such as the use of company-owned buildings, funds, or the time of the owners. Implicit costs do not require a cash payment and are not recorded in the accounting statement, which only documents the firm's revenues, explicit costs, and accounting profit.

---

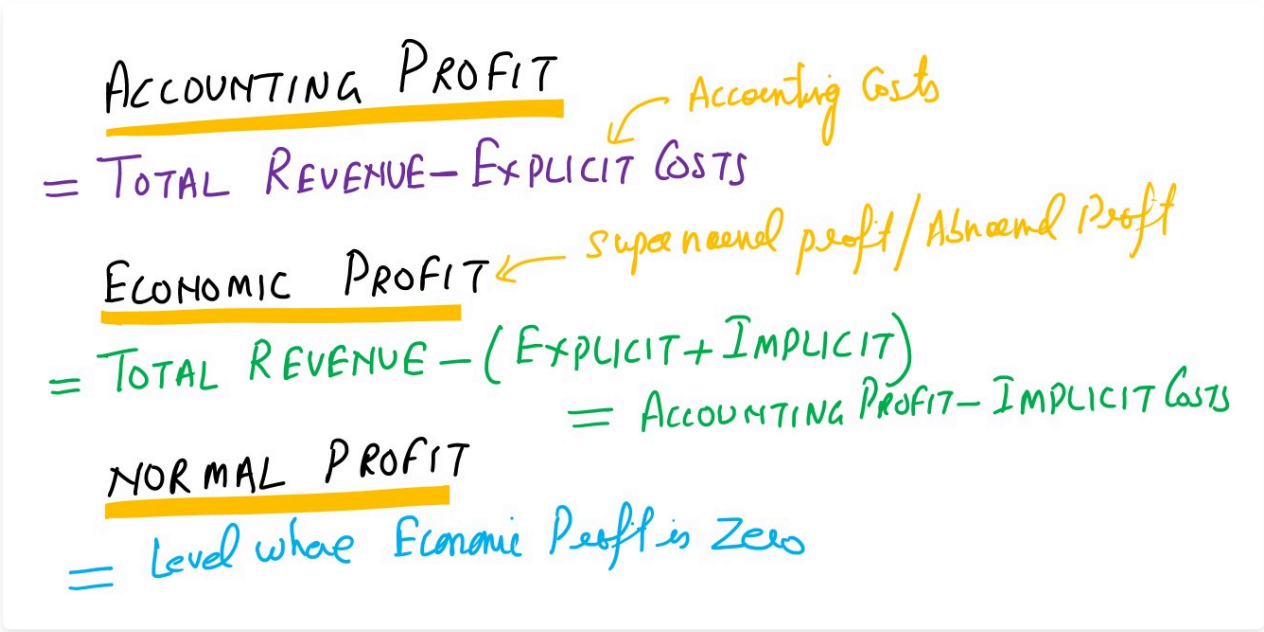
1. Introduction

A sunk cost is a cost that has already been incurred and cannot be recovered or changed by any present or future decision. In other words, it's a past expense that is irrelevant to current and future decision-making because the money spent cannot be retrieved. Sunk costs should not influence decisions, as they are independent of any potential benefits or losses associated with a current or future course of action.

---

1. Introduction

There are various ways to measure profits, each taking into account different types of costs. We will discuss three profit measures: accounting profit, economic profit, and normal profit, and their relationship with each other.



Accounting Profit

Accounting profit is the difference between a company's total revenue and its explicit costs. Explicit costs are those that are directly incurred by the business and can be easily measured, such as wages, rent, and materials. Accounting profit does not take into account implicit costs, which are the opportunity costs of using the resources of the business owner, such as their time or personal capital. The Explicit costs area also called Accounting Costs.

Accounting Profit = Total Revenue - Explicit Costs.

Economic Profit

Economic profit, on the other hand, takes into account both explicit and implicit costs. It is the difference between a company's total revenue and all of its costs, including both explicit and implicit costs. Economic profit is a more accurate measure of a company's profitability because it accounts for all of the costs associated with the business.

Economic Profit = Total Revenue - Total Economic Costs
Economic Profit = Total Revenue - (Explicit Costs + Implicit Costs)
Economic Profit = Accounting Profit - Implicit Opportunity Costs

The Economic Profit is also called *Abnormal Profit, Supernormal Profit* or *Pure Profit*.

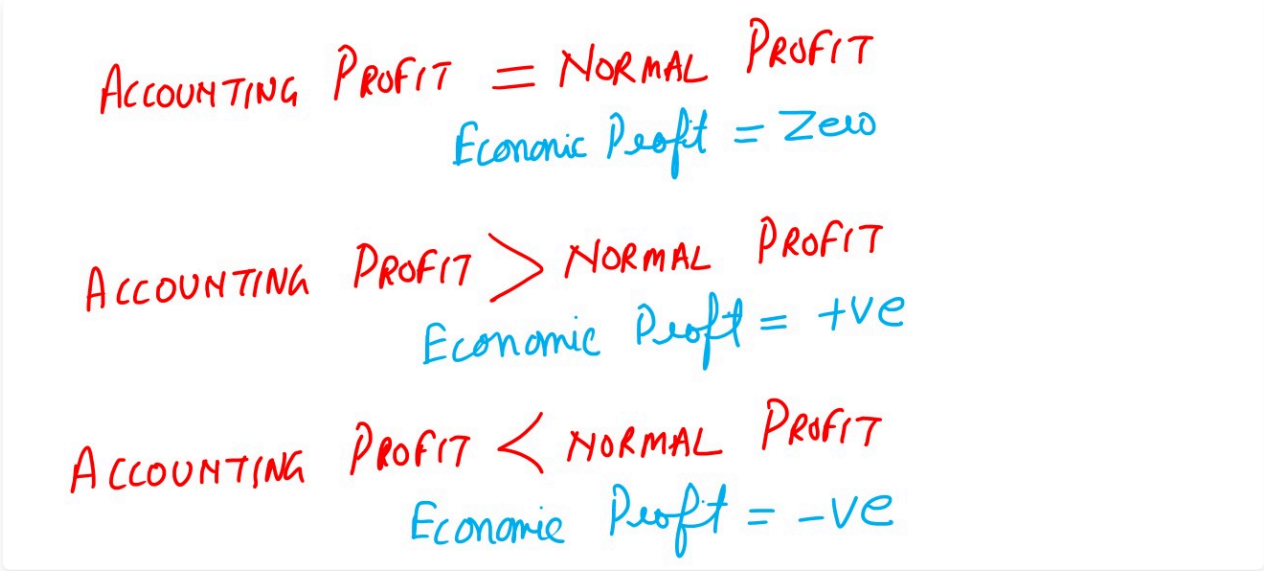
Let's take an example to illustrate these concepts. Suppose a small business owner named Sameer runs a landscaping company. In a year, he earns Rs 2,00,000 in revenue from his services. His explicit costs are Rs 80,000, which includes his employee salaries, materials, and rent. To calculate accounting profit, we subtract explicit costs from revenue, so his accounting profit is Rs 1,20,000.

However, Sameer also spends a significant amount of his time managing the business, and he could have earned Rs 50,000 if he had worked for another company instead. Additionally, he invested Rs 30,000 of his personal capital to purchase equipment for the business. If we take these implicit costs into account, Sameer's economic profit is Rs 40,000 (Rs 200,000 - Rs 80,000 - Rs 50,000 - Rs 30,000).

Normal Profit

Normal profit is the minimum amount of profit necessary to keep a business in operation. Normal profit is used to determine whether a business is earning enough profit to stay in operation. If a company is earning only normal profit, it is covering all of its costs (both explicit and implicit), but it is not earning any additional profit. It is also called Zero Economic Profit.

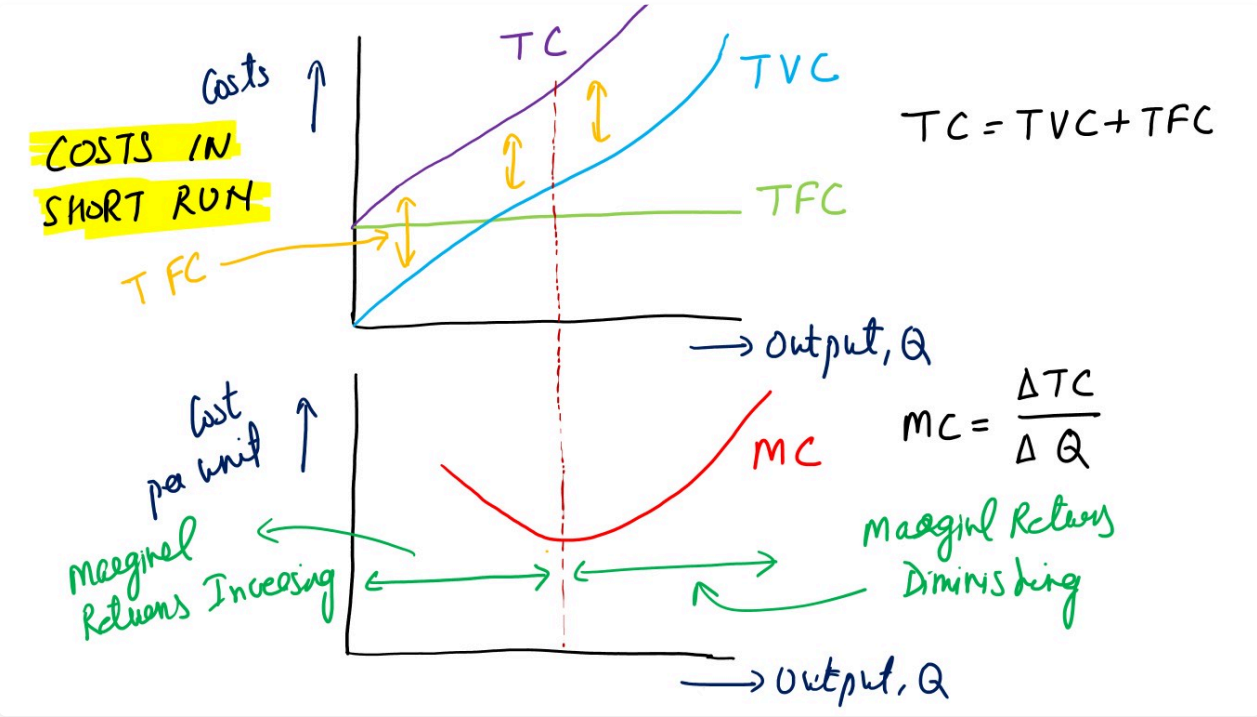
Normal Profit is Accounting Profit, which is required to cover Implicit Opportunity Costs.



If Accounting Profit = Normal Profit, then Economic Profit will be Zero.
If Accounting Profit > Normal Profit, then Economic Profit will be Positive.
If Accounting Profit < Normal Profit, then Economic Profit will be Negative.

---

2. Costs in Short Run



The cost of producing something can be divided into fixed and variable costs. **Fixed costs** are expenses a business has to pay regardless of how much they produce. For example, a business has to pay for property taxes, insurance, and equipment even if they don't produce anything. **Variable costs**, on the other hand, are expenses that increase as production increases, such as labour costs. If a business hires more workers, their variable costs increase. So, the total cost of producing something is the sum of fixed costs and variable costs.

In the short run, a business has fixed costs that do not vary with the level of production and variable costs that change as production levels change. Let us consider how the cost of production varies as output varies, in case of short run (at least one resource is fixed).

The behaviour of the various costs can be illustrated using the example of a small bakery that produces cupcakes, as shown in the following table.

Units of Production	Variable Cost	Total Cost	Marginal Cost	Average Variable Cost	Average Total Cost
0	0	50	-	-	50
1	10	60	60	10	60
2	20	70	10	10	35
3	25	80	10	8.33	26.67
4	30	95	15	7.5	23.75
5	40	120	25	8	24

The fixed cost is Rs 50. The figure shows cost curves for the data in the table.

Because fixed cost does not vary with output, the fixed cost curve is a horizontal line, parallel to x axis. Variable cost is zero when output is zero, so the variable cost curve starts from the origin. The total cost curve sums the fixed cost curve and the variable cost curve. Because a constant fixed cost is added to variable cost, the total cost curve is just the variable cost curve shifted vertically by the amount of fixed cost.

Total Cost is the sum of the fixed and variable costs of production. As the bakery produces more cupcakes, the total cost (TC) increases. At zero production, the total cost is equal to the fixed cost of Rs 50. As production increases, the total cost increases, but at a decreasing rate due to the law of diminishing returns.

Marginal Cost is the additional cost incurred by producing one additional unit of output. Marginal cost (MC) initially decreases with increasing production, reaches a minimum point (in this case at 2 units produced), and then increases as production continues. Marginal Cost initially decreases due increasing returns, but eventually increases due to diminishing returns.

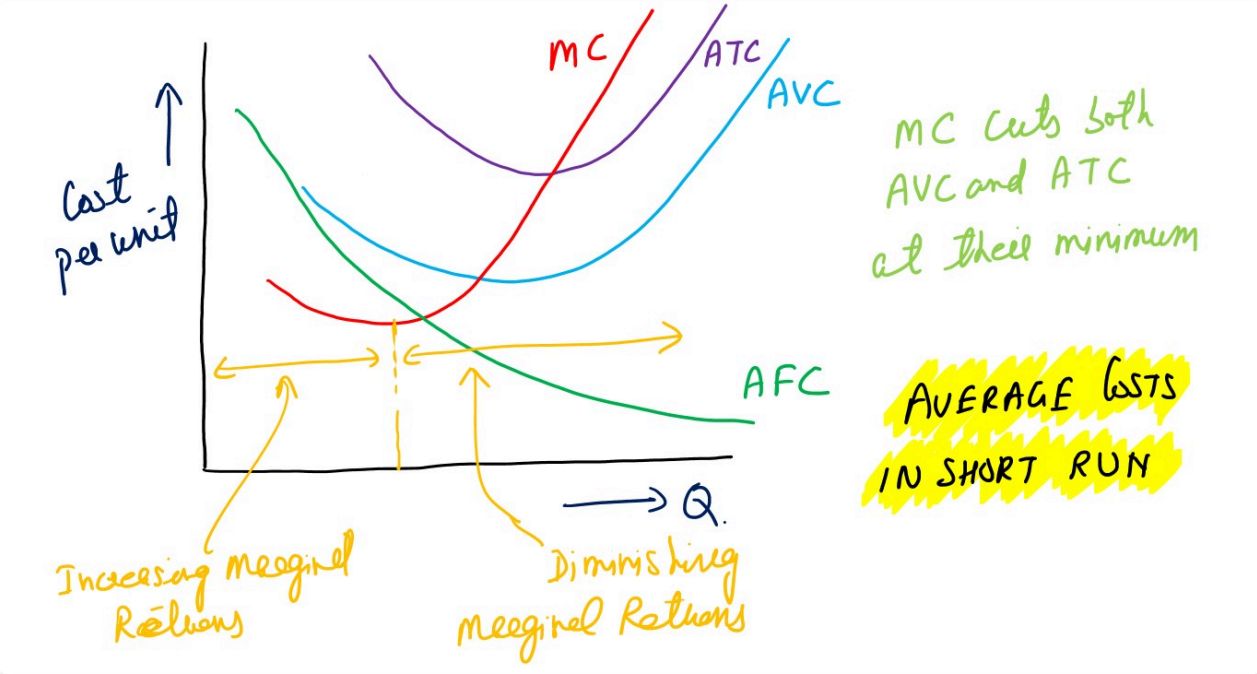
The MC curve is nothing but the slope of the TC curve at each rate of output. The TC curve can be divided into two sections, based on what happens to marginal cost:

1. Because of increasing marginal returns from labour, MC at first declines, so TC initially increases by successively smaller amounts and the TC curve becomes less steep.
2. Because of diminishing marginal returns from labour, MC starts increasing after the a few units of output, leading to a steeper TC curve.

Notice that the total cost curve has a backward S shape, the result of combining the two sections discussed above.

---

2. Costs in Short Run



The **Average Total Cost** (ATC) is the total cost per unit of output. The **Average Total Cost** (ATC) is also called **Average Cost** (AC). It initially decreases with increasing production, reaches a minimum point (in this case at 4 units produced), and then increases as production continues.

Initially there is increasing marginal return and MC is declining. As long as MC is below AC, the AC will keep declining. Then MC will reach at its lowest point and start rising (because of diminishing returns). Even at this stage, AC will keep declining, as long as MC is below AC. But when MC will cut AC from below, then AC will start rising. Thus, the point of intersection of AC and MC gives us lowest point of AC.

Average Total Cost (called ATC or AC) is the sum of average fixed cost (AFC) and average variable cost (AVC).

As the rate of output increases, **Average Fixed Cost** (AFC) progressively decreases because the fixed cost is distributed among a larger number of units. AFC consistently declines, eventually approaching the X-axis in an asymptotic fashion.

The **Average Variable Cost** (AVC) is the variable cost per unit of output. AVC curves behaves similar to AC curve. Initially AVC decreases with increasing production and reaches a minimum point (where MC meets AVC from below) before increasing again. Thus, the point of intersection of AVC and MC gives us lowest point of AVC.

Thus, the rising MC curve intersects both the AVC curve and the AC curve where these average curves are at their minimum. This occurs because the marginal pulls down the average where the marginal is below the average and pulls up the average where the marginal is above the average.

The most optimum level of production is given by that quantity of production, where the MC cuts AC from below. This point is called **Economic Capacity**.

Thus, we see that the law of diminishing returns dictates the shape of short run curves.

Illustration

A firm's total cost function is, $TC = 12 + 60Q - 15Q^2 + Q^3$

Suppose that the firm produces 10 units of output. Calculate total fixed cost (TFC), total variable cost (TVC), average total cost (ATC), average fixed cost (AFC), average variable cost (AVC), and marginal cost (MC).

Solution:

We can see from the formula of TC that 12 is TFC (does not depend on Q).

Thus, the TVC is $= 60Q - 15Q^2 + Q^3 = 60(10) - 15(10)^2 + (10)^3 = 100$

ATC (obtained from dividing TC by Q) = 12Q-1 + 60 -15Q +Q^2= \(\frac{12}{10} \) + 60 - 15(0) + $(10)^2$= 11.2

AFC (obtained from dividing TFC by Q) = $12Q^{-1}$ = \(\frac{12}{10} \) = 1.2

AVC (obtained from dividing TVC by Q) = 60 – 15Q + Q^2 = 60 – 15(10) + $(10)^2$ = 10

MC (obtained from differentiating TC w.r.t Q)

\(= \frac{dTC}{dQ} \) = 60 – 30Q + $3Q^2$ = 60 – 30(10) + $3(10)^2$ =60

---

3. Costs in Long Run

The focus of the analysis thus far has been on how a firm's costs change in the short run when output rates increase, while the size of the firm remains the same. In the long run, however, a firm can adjust all inputs that it controls, and as such, there are no fixed costs. The long run should not be viewed as a series of short runs, but rather as a planning horizon. In the long run, the firm has flexibility in selecting input combinations, but once the size of the plant has been determined and constructed, the firm is locked into fixed costs and is operating in the short run. While **firms plan for the long run, they produce in the short run**.

To provide an example, imagine a bakery that has a fixed number of ovens, mixers, and employees. In the short run, if demand for the bakery's goods increases, the bakery can hire more employees and increase production, but it cannot purchase additional ovens or mixers. In the long run, however, the bakery can expand its operations by purchasing additional equipment and hiring more employees, allowing it to produce more goods without incurring fixed costs.

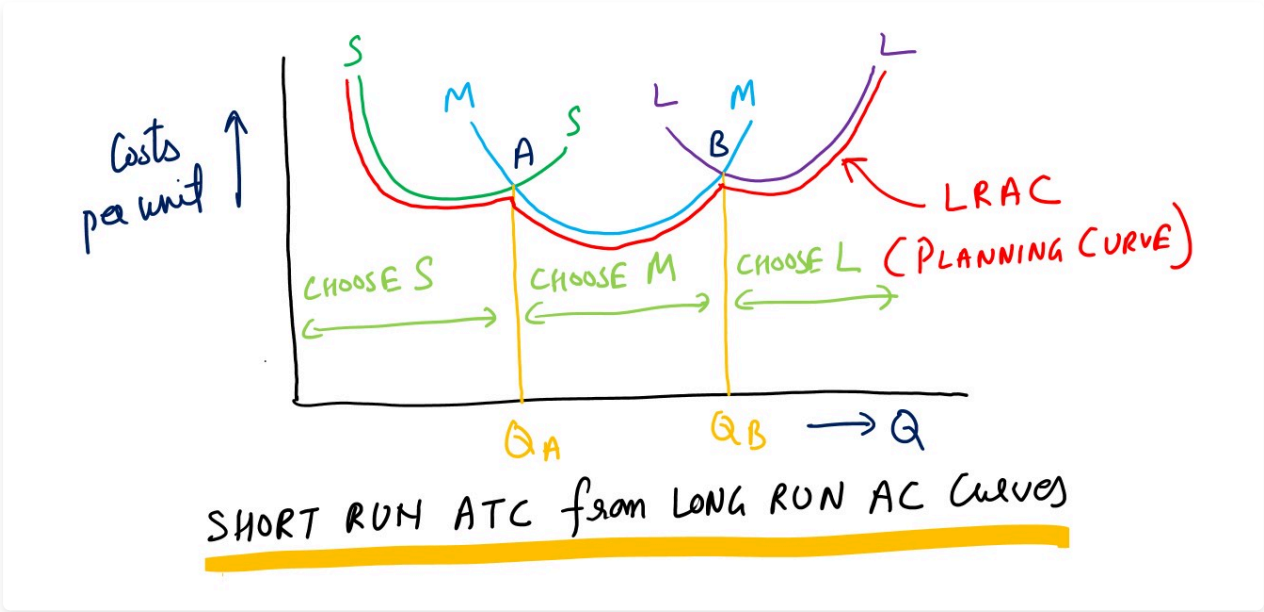
In the short run, we study increasing and diminishing returns from the variable resource. In the long run, it is economies and diseconomies of scale.

---

3. Costs in Long Run

Suppose a firm must choose from among three possible plant sizes: small, medium, and large. The figure shows this simple case. The average cost (AC) curves for the three sizes are SS, MM, and LL. The long-run average-cost (LRAC) curve is SABL. We have to determine, which size should the firm build to minimize average cost. The appropriate size, or scale, for the firm depends on how much the firm wants to produce.

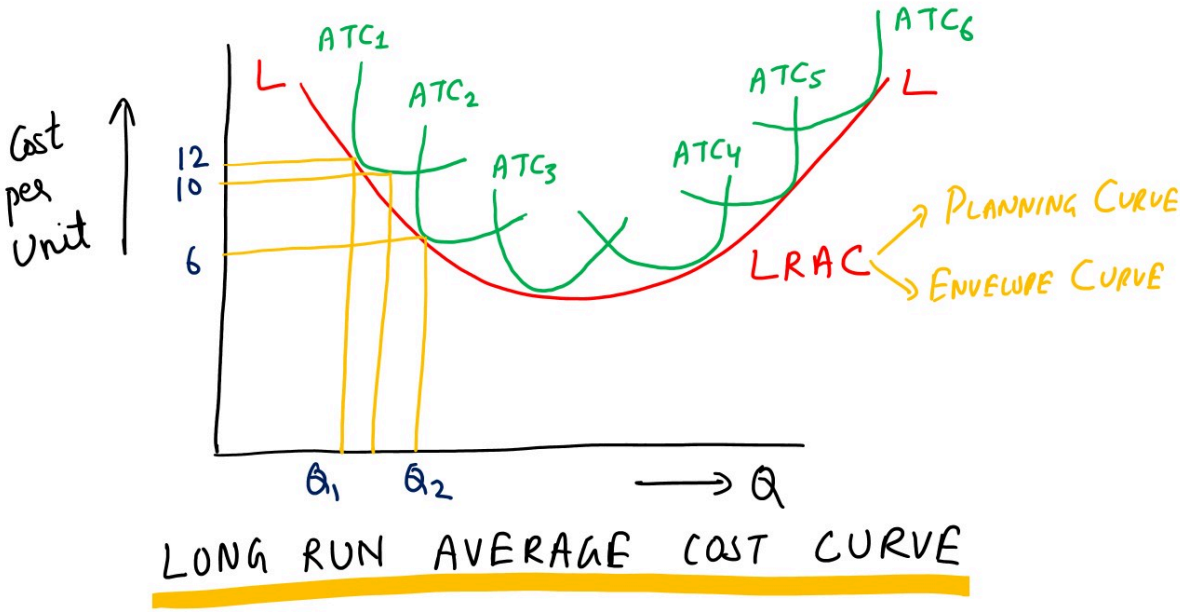
For any output less than Q_A, AC is lowest when the plant is small. For output between Q_A and Q_B, AC is lowest for the plant of medium size. And for output that exceeds Q_B, AC is lowest when the plant is large.



The LRAC curve, sometimes called the **firm's planning curve**, connects portions of the three short-run average cost curves that are lowest for each output rate.

Now suppose there are many possible plant sizes. The figure presents a sample of short run cost curves ATC_1, ATC_2, ATC_3 and so on). The LRAC curve, shown by LL, is formed by connecting the points on the various short-run average cost curves that represent the lowest per-unit cost for each rate of output. Each of the short-run average cost curves is tangent to the LRAC curve, or planning curve. If we could display enough short-run cost curves, we would have a different plant size for each rate of output. These points of tangency represent the least-cost way of producing each particular rate of output, given the technology and resource prices.

For example, the short-run average total cost curve ATC_1 is tangent to the long run average cost curve at point X, where Rs 12 is the lowest average cost of producing output Q_1. Note, however, that other output rates along ATC_1 have a lower average cost. For example, the average cost of producing Q_2 is only Rs 10, as identified at point Y. Point Y depicts the lowest average cost along ATC_1. So, while the point of tangency reflects the least-cost way of producing a particular rate of output, that tangency point does not reflect the minimum average cost for this particular plant size.



If the firm decides to produce Q_2, which size plant should it choose to minimize the average cost of production? Output rate Q_2 could be produced at point Y, which represents the minimum average cost along ATC_1.

But average cost is lower with a larger plant. With the plant size associated with ATC_2, the average cost of producing Q_2 would be minimized at Rs 6 per unit at point Z. Each point of tangency between a short-run average cost curve and the LRAC curve represents the least-cost way of producing that particular rate of output.

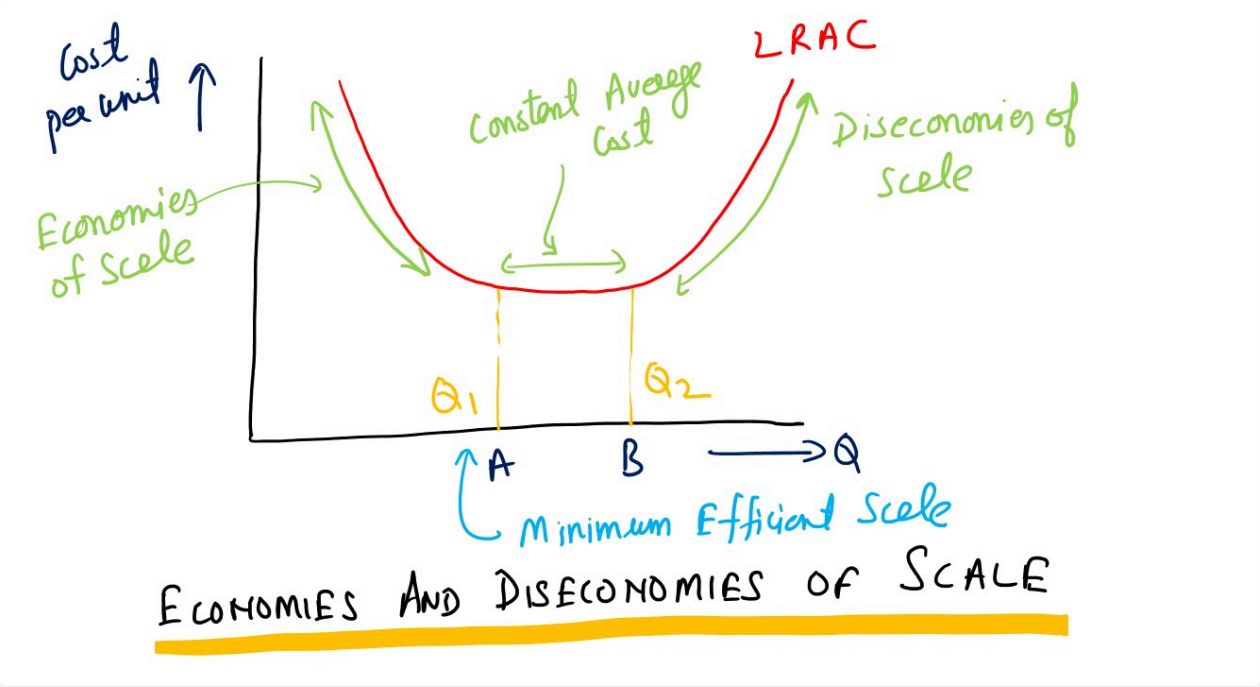
Thus, with many possible plant sizes, the LRAC curve is the **envelope** of portions of the short-run average cost curves. Each short-run curve is tangent to the LRAC curve. Each point of tangency represents the least-cost way of producing that level of output.

As depicted in the diagram, the LRAC curve exhibits a U shape. It's important to note that both the AVC and AC curves in the short run also displayed a U shape, but the underlying cause differed. In the short run, the U shape of AC and AVC resulted from the law of diminishing marginal returns, whereas the U shape of the LRAC curve is attributed to the presence of Economies and Diseconomies of Scale.

The output level that corresponds to minimum LRAC curve is commonly referred to as the **Minimum Efficient Scale (MES)**. The MES is the level of output that corresponds to the lowest per-unit cost of production in the long run.

---

4. Economies and Diseconomies of scale



Economies of scale refer to the cost advantages that firms can achieve by expanding their production output. As a firm expands, it can often take advantage of various cost-saving measures, such as using larger and more specialized machines and assigning workers to more specialized tasks. This results in lower long-run average costs as output expands.

For example, a small restaurant with a household-size kitchen may produce meals at a lower average cost than a larger restaurant like McDonald's at low rates of output. However, if production in the smaller kitchen increases beyond a certain point, a larger-scale kitchen like McDonald's would be able to produce at a lower average cost due to the economies of scale.

Diseconomies of scale, on the other hand, occur when a firm expands its plant size to the point where the costs of coordination and communication outweigh the benefits of larger-scale production. As the amount and variety of resources employed increase, coordinating all these inputs becomes more challenging. Additional layers of management may be needed to monitor production, leading to a thicket of bureaucracy that can reduce efficiency and increase average cost. This often results in higher long-run average costs as output expands. For example, rumors may become a primary source of information in very large organizations, leading to reduced efficiency and increased average cost.

Sometimes, a firm may experience neither economies of scale nor diseconomies of scale over some range of output, resulting in **constant long-run average cost**. This means that average cost neither increases nor decreases with changes in firm size. This could occur because economies and diseconomies of scale exist simultaneously in the firm but have offsetting effects. For instance, a firm may experience cost-saving measures due to economies of scale in some areas, but also face coordination challenges due to diseconomies of scale in other areas.

As an example, consider a movie theater. Economies of scale could be achieved by investing in larger screens and better sound systems, allowing for greater specialization of labour and more efficient use of space. However, diseconomies of scale could arise from the need to manage a larger workforce and coordinate more complex operations. For instance, managing a larger number of screens and showtimes could lead to communication challenges and reduce efficiency. If the firm experiences both economies and diseconomies of scale but with offsetting effects, it may achieve constant long-run average cost over a range of output. The firm could find the minimum efficient scale (MES), which is the lowest rate of output at which LRAC is at a minimum.

---

1. Market Structures

The **market structure** refers to the critical characteristics of a market that impact business decision-making. These characteristics include the number of suppliers (high or low concentration), product homogeneity (whether offerings are identical or differentiated), barriers to entry (ease or difficulty for new entrants), and competitive strategies (price-based competition, advertising, or product differentiation). By understanding these aspects, managers can make informed decisions on production levels and pricing strategies.

It may be noted that the terms "industry" and "market" are used interchangeably in our discussion. An industry comprises all firms supplying products to a specific market, such as the automobile, footwear, or agricultural commodities markets.

The nature of competition that a product faces is a crucial determinant of its price. Competition reflects the product's position in the market compared to its rivals, taking into account the number of buyers, sellers, degree of product differentiation, and entry-exit barriers. Various combinations of these characteristics result in diverse market structures, including perfect competition, monopolistic competition, monopoly, oligopoly, and monopsony. These categories are based on the level of competition, with perfect competition exhibiting the highest competition and monopoly the lowest competition.

---

2. Perfect Competition

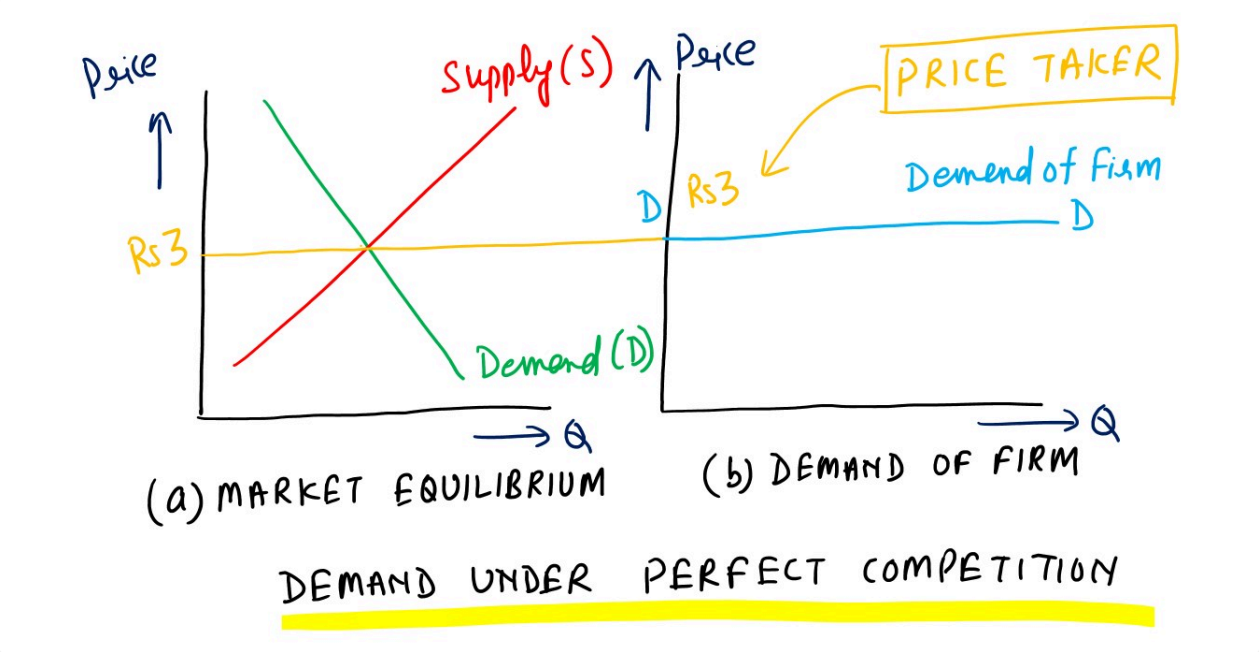
We commence our exploration with the perfect competition, which can be considered one of the most fundamental market structures. A perfectly competitive market exhibits the following key characteristics:

1. **Abundance of Buyers and Sellers:** In such a market, there are a multitude of buyers and sellers, to the extent that each participant engages in transactions that represent only a minuscule portion of the overall market activity.
2. **Homogeneous Products:** Firms operating within this structure offer standardized commodities. These products, are indistinguishable from those offered by other producers. In essence, there is no product differentiation among competitors.
3. **Full Information:** Participants in this market possess complete and accurate knowledge regarding the prevailing prices and the availability of all resources and products. This transparency ensures that buyers and sellers make well-informed decisions.
4. **Ease of Mobility:** Firms and resources in a perfectly competitive environment can easily enter or exit the industry over time. They encounter minimal obstacles such as patents, licenses, high capital requirements, or lack of awareness about available technologies when making these decisions. This fluidity facilitates competition and market efficiency.

---

3. Demand under Perfect Competition

Consider a market for a good, such as apples, with numerous apple orchards operating within it. Given the vast number of apple orchards, each one contributes only a minuscule portion of the overall market output.



As illustrated in the figure, the market price of apples at Rs 3 per Kg is established in panel (a) by the intersection of the market demand curve (D) and the market supply curve (S). At this market price of Rs 3 per Kg, any apple orchard can sell its entire output.

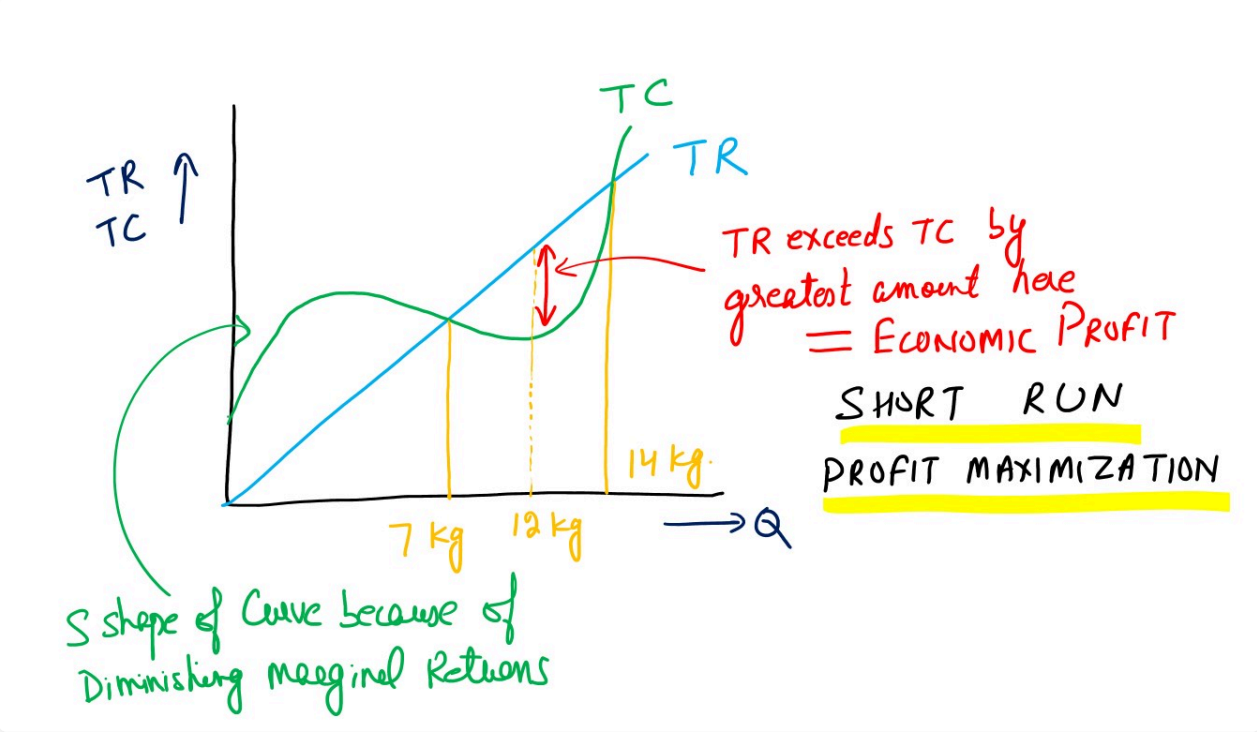
Since each orchard is relatively small compared to the entire market, it has no influence on the market price. Additionally, as all orchards produce an identical product—apples in this case—any orchard that charges more than the market price will not make any sales. For example, an orchard charging Rs 3.10 per Kg would find no buyers.

Consequently, the demand curve faced by an individual orchard, denoted as curve DD in panel (b), is a horizontal line at the market price of Rs 3 per Kg. This implies that each orchard faces a horizontal or perfectly elastic demand curve. A firm in perfect competition is considered a **price taker** because it must accept the market price—essentially, it's a "take it or leave it" scenario.

---

4. Short Run Profit Maximization

The firm is said to be operating in the short run, when there is a presence of fixed cost (at least one resource is fixed).



The firm maximizes economic profit by finding the quantity at which total revenue (TR) exceeds total cost (TC) by the greatest amount. The firm's TR is simply its output (Q) times the price per unit (P). In case of perfect competition, the market price of Apple (Rs 3 per Kg), does not vary with the firm's output. As output increases by 1 Kg, TR increases by Rs 3, so the firm's total revenue curve is a straight line emanating from the origin, with a slope of 3.

The short run TC curve has the backward S shape, showing increasing and then diminishing marginal returns from the variable resource. Total cost always increases as more output is produced.

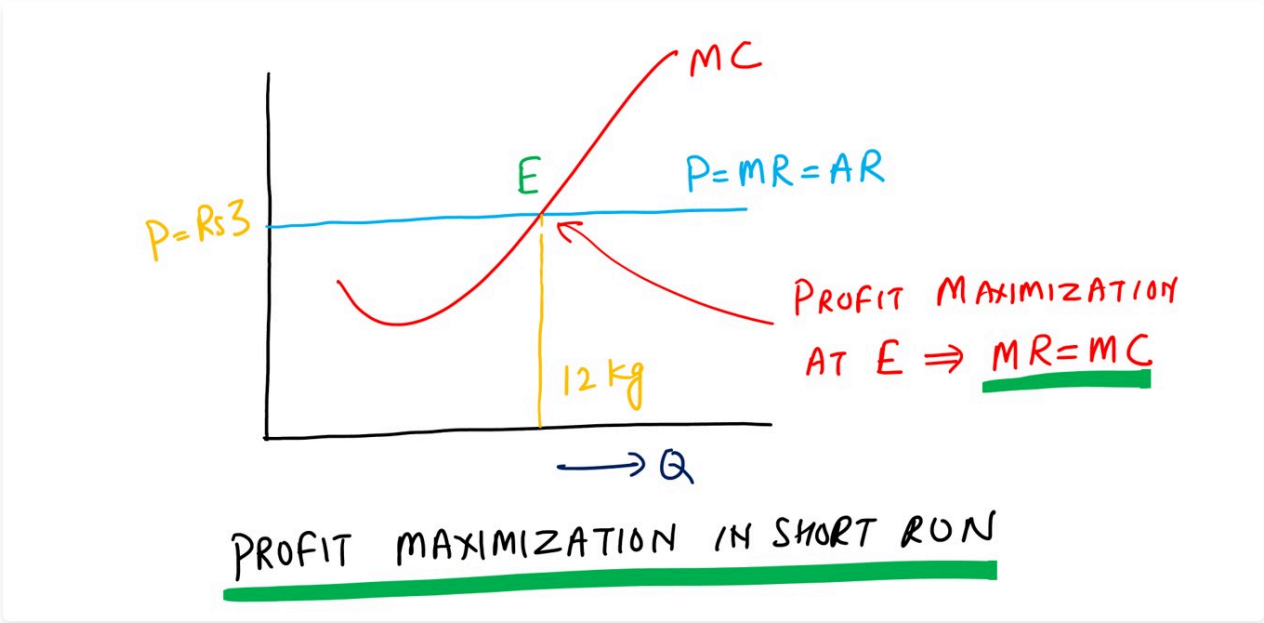
As we can see in the diagram, TR exceeds TC when 7 to 14 Kg apples are produced, so the firm earns an economic profit at those output rates. The Economic profit is maximized, when the firm produces 12 Kg of apples.

Subtracting total cost from total revenue is one way to find the profit-maximizing output. For output less than 7 Kg and greater than 14 Kg, TC exceeds TR. The **economic loss** is measured by the vertical distance between the two curves. Between 7 and 14 Kg, TR exceeds TC. The **economic profit**, again, is measured by the distance between the two curves.

---

4. Short Run Profit Maximization

Another way to find the profit-maximizing rate of output is to focus on marginal revenue (MR) and marginal cost (MC).



In perfect competition, each firm is a price taker, so selling one more unit increases TR by the market price (P). Thus, in perfect competition, MR is equal to the market price (P). The MR is a horizontal line at the market price P of Rs 3.

P = MR

The MC first declines, reflecting increasing marginal returns in the short run as more of the variable resource is employed. MC then increases, reflecting diminishing marginal returns from the variable resource.

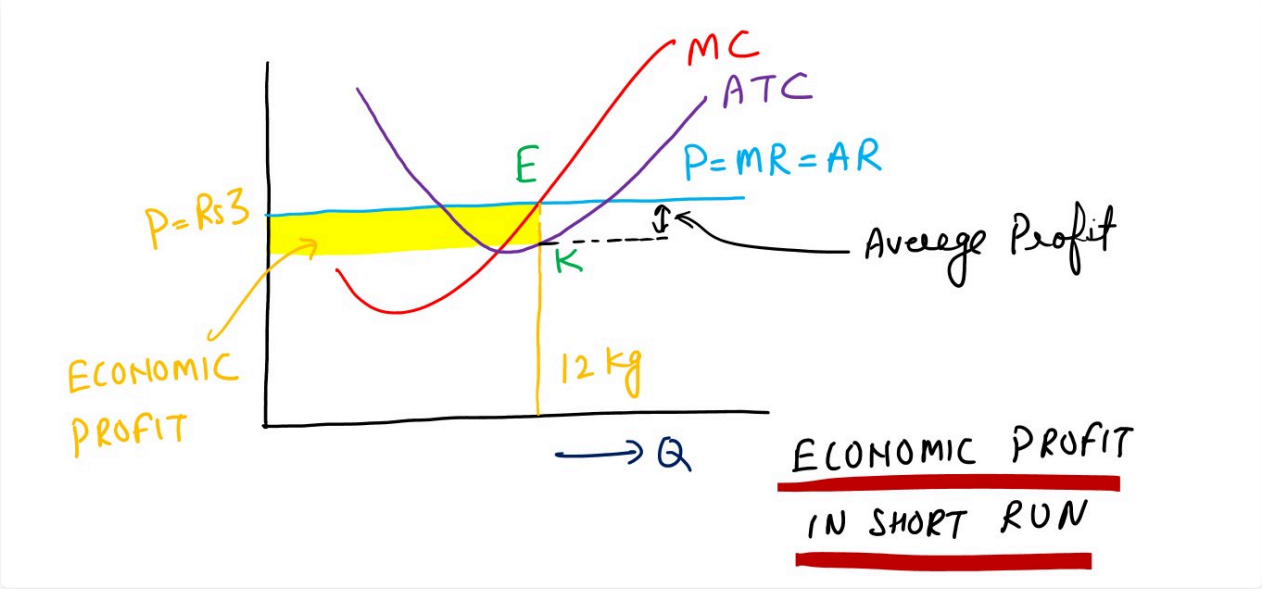
The firm will increase production as long as each additional unit adds more to TR than to TC—that is, as long as MR exceeds MC.

In our figure, MR exceeds MC for each of the first 12 Kg of apples. Producing 13 Kg or 11 Kg would reduce economic profit. The firm, as a profit maximizer, will limit output to 12 Kg.

Thus, the firm will expand output as long as MR exceeds MC and will stop expanding before MC exceeds MR. A shorthand expression for this approach is the golden rule of profit maximization, which says that a profit maximizing firm produces the quantity where MR equals MC. This is represented by point E in our figure.

We have understood that in perfect competition, MR equals the market price (P), which is also the perfectly competitive firm's demand curve. Because the perfectly competitive firm can sell any quantity for the same price per unit, the average revenue (AR) will also be equal to price (P). Thus, now we have,

Market Price (P) = Marginal Revenue (MR) = Average Revenue (AR) = Firm's demand curve

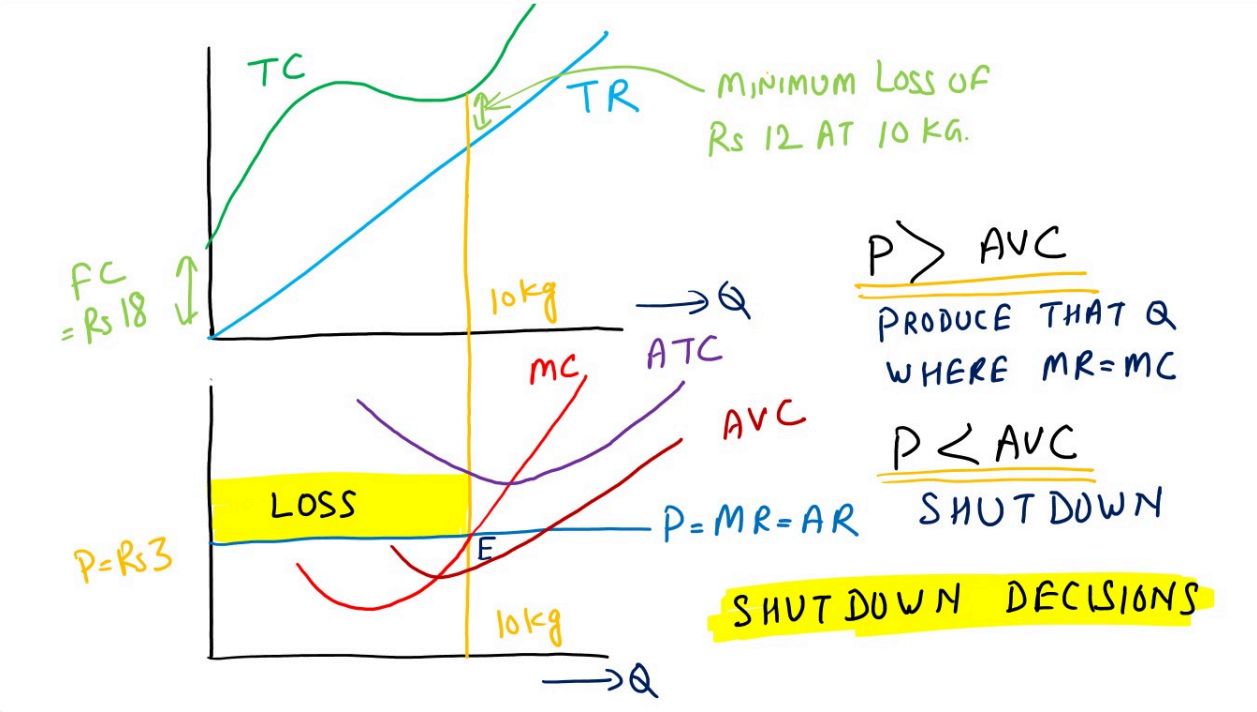


The Economic Profit is represented by the shaded rectangle. The height of that rectangle (represented by KE), equals the price (P) or AR of Rs 3 minus the average total cost (ATC), say Rs 2. Price minus ATC yields an average profit of Rs 1 per Kg. Total Profit, Rs 12, equals the average profit per Kg, Rs 1 (denoted by KE), times the 12 Kg produced.

To conclude, with the TC and TR curves, we measure economic profit by the vertical distance between the two curves. But with the per-unit curves, we measure economic profit by an area—that is, by multiplying the average profit of Rs 1 per Kg times the 12 Kg sold.

---

5. Minimizing Short-Run Losses- Shutdown decision



Till now, we have learnt that in a perfect competition, a firm cannot control the price of its product. Sometimes the price can be so low that the company loses money no matter how much it produces. The company can choose to keep producing at a loss or temporarily stop production. So, should a company continue to produce at a loss or stop production?

In the short term, the company has two types of costs: fixed costs (FC like taxes and insurance) that must be paid even if the company produces nothing, and variable costs (VC like labor) that depend on the amount of production. If the company stops production, it still has to pay its fixed costs.

But if it produces, the revenue can cover variable costs and a portion of fixed costs. Therefore, a company will produce instead of shutting down if the TR is more than the variable cost (TVC) of production. The excess revenue (TR-TVC) can go toward covering at least a portion of fixed cost.

As we can see in the figure, the total cost (TC) curve always exceeds total revenue (TR), the firm suffers a loss no matter how much is produced. The vertical distance between the two curves measures the loss at each quantity. If the firm produces nothing, the loss is the fixed cost of Rs 18. The vertical distance between the two curves is minimized at 10 Kg, which is a loss, but less than Rs 18.

We get the same result using marginal analysis.

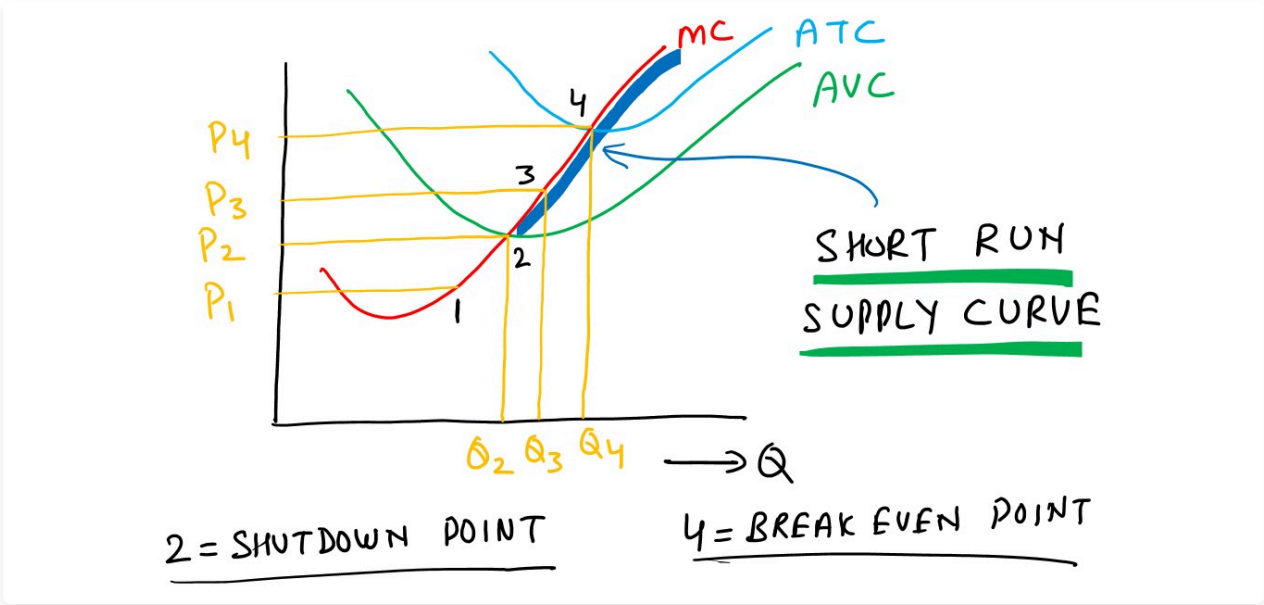
We can see in the figure that the marginal revenue (MR) equals marginal cost (MC) at point E. The loss is equal to output of 10 Kg multiplied by the difference between average total cost (ATC) and price (Rs 3). Because the price (P) exceeds average variable cost (AVC), the firm is better off continuing to produce in the short run, since revenue covers some fixed cost.

The bottom line is that the firm will produce rather than shut down if there is some rate of output where the price (P) at least covers average variable cost (AVC). But if the average variable cost (AVC) exceeds the price (P) at all rates of output, the firm will shut down.

Shutting down is not the same as going out of business. In the short run, even a firm that shuts down keeps its productive capacity intact—paying for rent, insurance, and property taxes, keeping water pipes from freezing in the winter, and so on. The short run is defined as a period during which some costs are fixed, so a firm cannot escape those costs in the short run, no matter what it does. Fixed cost is sunk cost in the short run, whether the firm produces or shuts down.

6. Short-run supply curve of firm

Till now, we learnt that, if average variable cost (AVC) exceeds price (P = MR) at all output rates, the firm will shut down in the short run. But if price (P) exceeds average variable cost (AVC), the firm will produce the quantity at which marginal revenue (MR) equals marginal cost (MC).



At point P_1

P < AVC and P < ATC

It is recommended to shutdown. The firm will incur loss.

At point P_2

P = AVC and P < ATC

The firm is indifferent to shutdown or continue producing. This is called **shutdown point.** The firm will incur loss.

At point P_3

P > AVC and P < ATC

It is recommended to continue producing. The firm will incur loss.

At point P_4

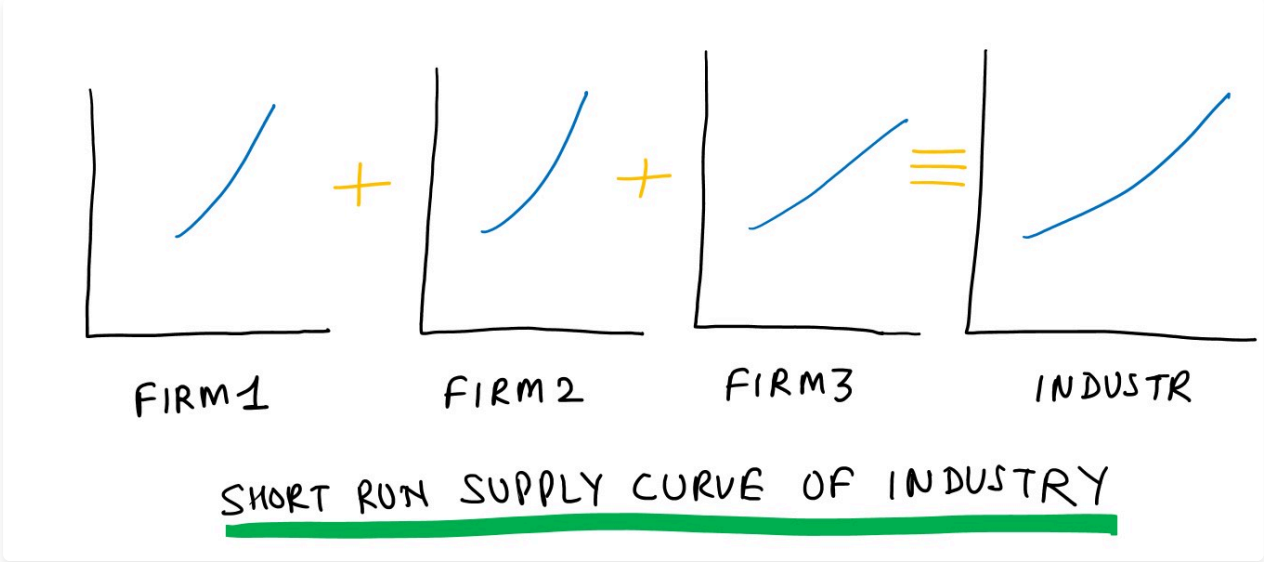
P > AVC and P = ATC

It is recommended to continue producing. This is called **breakeven point**. The firm will incur normal profit. (Economic Profit = 0)

At point P_5

P > AVC and P > ATC

It is recommended to continue producing. The firm will incur super normal profit (Economic Profit > 0)

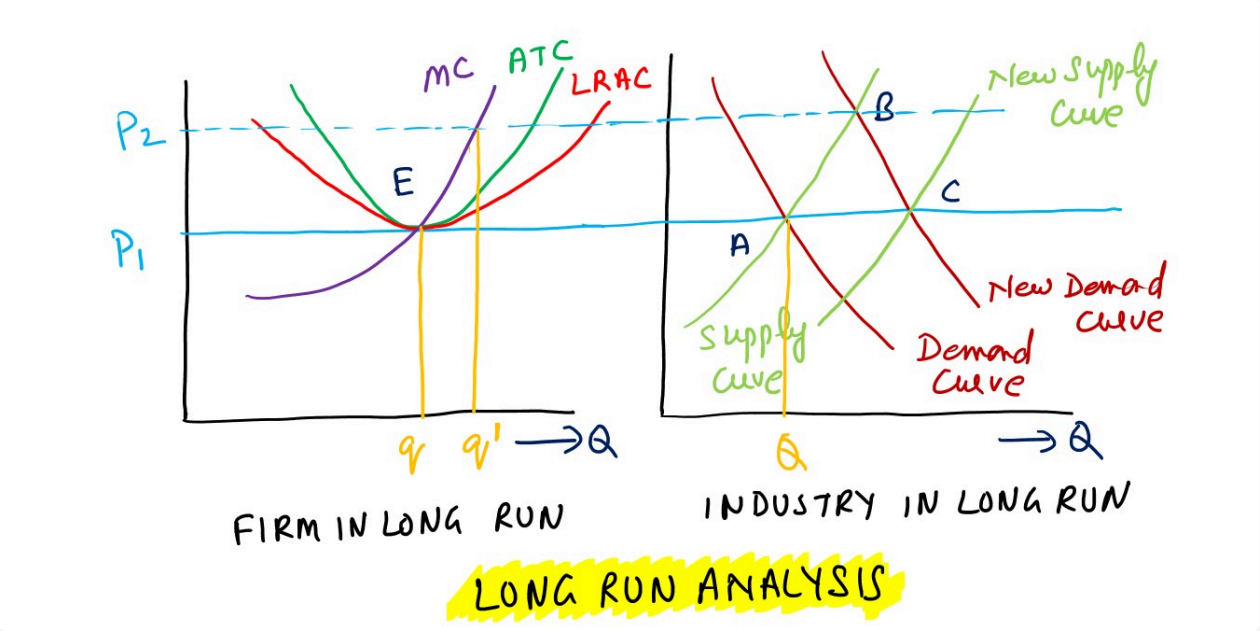


Thus, as long as the price (P) covers AVC, the firm will supply the quantity at which the upward-sloping marginal cost curve intersects the MR, or demand, curve. Thus, that portion of the firm's marginal cost (MC) curve that intersects and rises above the lowest point on its average variable cost (AVC) curve becomes the **short-run firm supply curve**. Thus, the short-run supply curve is the upward-sloping portion of the marginal cost curve, beginning at the shutdown point.

The **short-run industry supply curve** is the horizontal sum of all firms' short-run supply curves.

---

7. Perfect Competition in the Long Run



In the long run, there is no distinction between fixed and variable cost because all resources under the firm's control are variable.

Short-run economic profit will, in the long run, encourage new firms to enter the market and may prompt existing firms to get bigger. Economic profit will attract resources from industries where firms are losing money or earning only a normal profit. This expansion in the number and size of firms will shift the industry supply curve rightward in the long run, driving down the price. New firms will continue to enter a profitable industry and existing firms will continue to expand as long as economic profit is greater than zero. Entry and expansion will stop only when the resulting increase in supply drives down the price enough to erase economic profit.

On the other hand, a short-run loss will, in the long run, force some firms to leave the industry or to reduce their scale of operation. In the long run, departures and reductions in scale shift the market supply curve to the left, thereby increasing the market price until remaining firms just break even—that is, earn a normal profit.

In the long run, firms in perfect competition earn just a normal profit, which means zero economic profit.

The figure shows a firm and the market in long-run equilibrium. In the long run, market supply adjusts as firms enter or leave or change their size. This long-run adjustment continues until the market supply curve intersects the market demand curve at a price that corresponds to the lowest point on each firm's long-run average cost curve, or LRAC curve.

Because the long run is a period during which all resources under a firm's control can be varied, a firm in the long run will be forced by competition to adjust its scale until its average cost of production is minimized. A firm that fails to minimize cost will not survive in the long run.

At point E in the figure, the firm is in equilibrium, producing q units and earning just a normal profit. At point E, the price, marginal cost (MC), short-run average total cost (ATC), and long-run average cost (LRAC) are all equal. No firm in the market has any reason to change its output rate, and no outside firm has any incentive to enter this industry, because firms in this market are earning normal, but not economic, profit.

Suppose the market demand increases, as shown by the dotted New Demand Curve in the figure (Demand curve shifts to right). This leads to increase in equilibrium price from P_1 to P_2. The firm responds to the higher price by increasing output to q' and earns more economic profit. Economic profit attracts new firms to the industry in the long run. Since new firms are entering, the market supply shifts right (dotted supply curve), pushing the market price back down to P_1. The firm responds accordingly, by reducing production from q to q', leading to erosion of economic profit. The short-run adjustment is from point A to point B, but the long-run adjustment is from point A to point C.

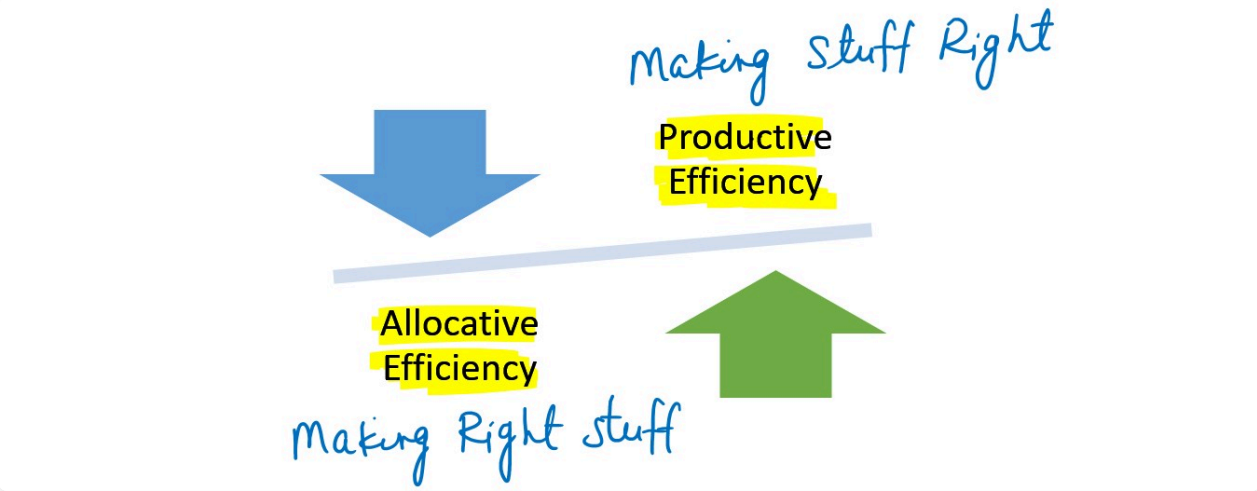
Thus, we learn that, new firms are attracted to the industry by short-run economic profits resulting from the increase in demand. But this new entry increases market supply, pushing the price down until economic profit disappears.

On the other hand, a short-run loss, if it continues, will in the long run force some firms out of business. As firms exit, market supply decreases (leftward shift of market supply curve), so the price increases. Firms continue to leave until the price is back to equilibrium point and remaining firms once again earn a normal profit (economic profit is zero).

---

8. Perfect Competition and Market Efficiency

Perfect competition guarantees both productive efficiency and allocative efficiency in the long run.



Productive Efficiency: Making Stuff Right

Productive efficiency occurs when a firm produces at the minimum point on its long-run average cost (LRAC) curve. In perfect competition, the market price equals the minimum average cost, ensuring that each firm produces at the minimum point on its long-run average cost curve. If firms do not reach the minimum long-run average cost, they must adjust their scale or leave the industry to avoid continued losses. Thus, perfect competition produces output at the minimum average cost in the long run.

For example, consider the agricultural industry where farmers produce wheat. In a perfectly competitive market, farmers must produce wheat at the lowest possible cost to compete with other farmers. As a result, they will use the most efficient methods of production to ensure that their costs are minimized. In the long run, each farmer will produce at the minimum point on their long-run average cost curve, leading to productive efficiency.

Allocative Efficiency: Making the Right Stuff

Allocative efficiency occurs when firms produce the output that is most valued by consumers. In perfect competition, firms produce goods efficiently and produce the right goods that consumers want. The market demand and supply curves determine the equilibrium price and quantity, which ensures that the marginal benefit consumers derive from the last unit consumed is equal to the opportunity cost of the resources employed to produce that unit.

When the marginal benefit that consumers derive from a good, equals the marginal cost of producing that good, that market is said to be allocatively efficient.

Marginal Benefit = Marginal Cost

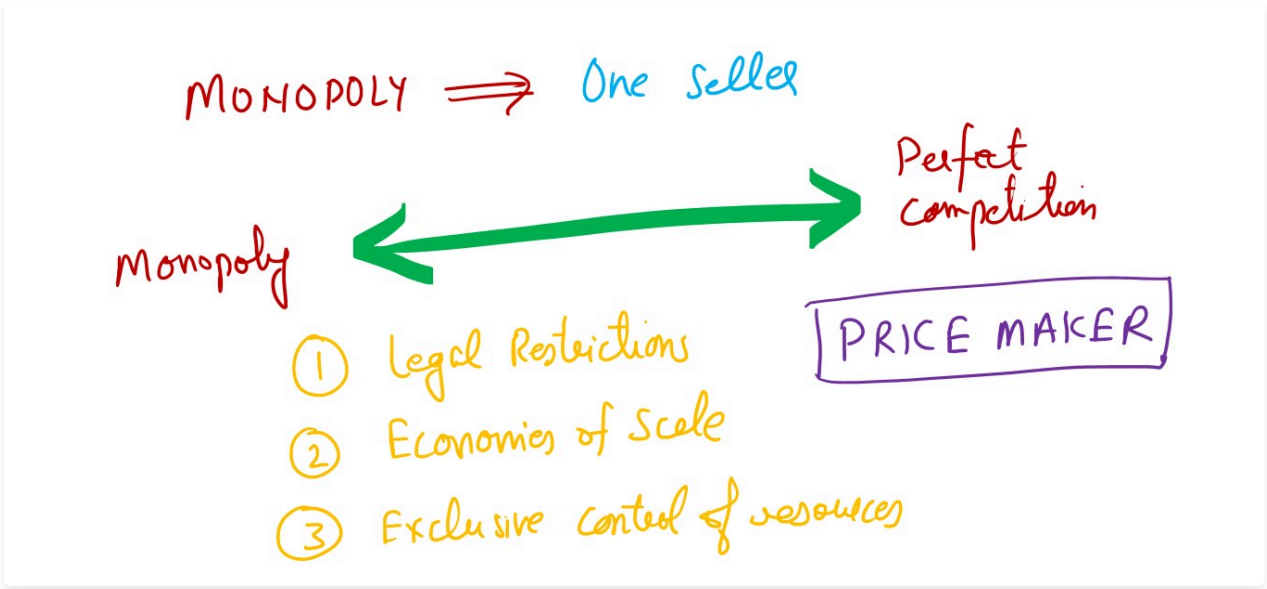
For example, consider the market for smartphones. In a perfectly competitive market, firms must produce smartphones that are in demand and provide the most value to consumers. Consumers will purchase the smartphone that provides the most utility for them at the lowest possible price. In the long run, firms that produce smartphones that are not in demand will exit the market, leaving only the firms that produce the most demanded smartphones. Thus, perfect competition ensures that firms produce the output that is most valued by consumers, leading to allocative efficiency.

Productive and allocative efficiency in the short run occurs at equilibrium point, which also is the combination of price and quantity that maximizes the sum of consumer surplus and producer surplus, thus maximizing social welfare. Social welfare is the overall well-being of people in the economy. Even though marginal cost equals marginal benefit for the final unit produced and consumed, both producers and consumers usually derive a surplus.

---

1. Monopoly

Monopoly is a market structure characterized by the presence of a single firm that is the sole supplier of a product with no close substitutes. In a monopolized market, there are significant barriers to entry that prevent new firms from entering the industry. These barriers may be legal restrictions, economies of scale, or the monopolist's control of an essential resource.



Let us discuss these features one by one.

---

1. Monopoly

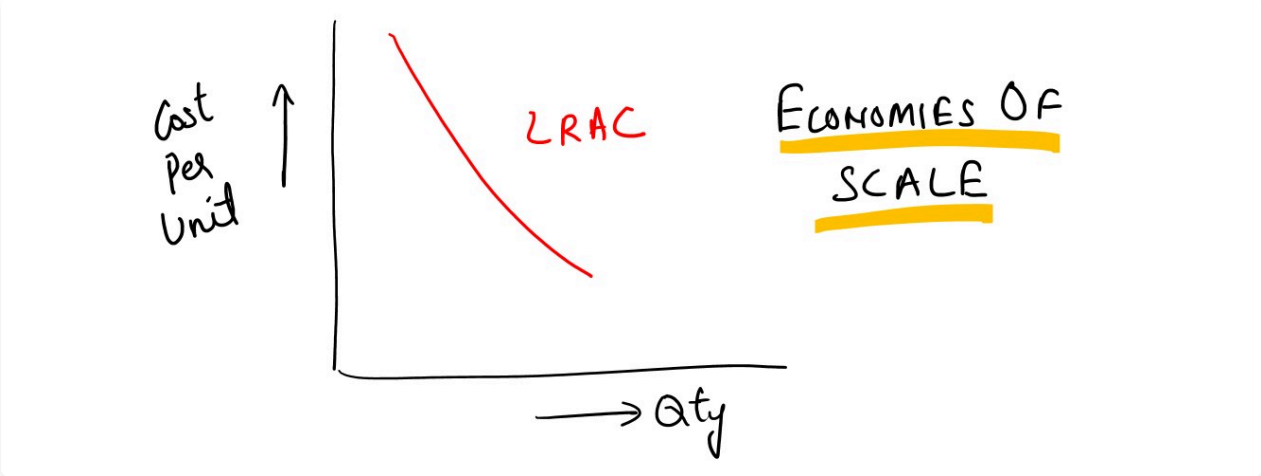
One type of barrier to entry is legal restrictions, which include patents, invention incentives, and licenses. A patent is a legal protection granted to an inventor that gives the inventor exclusive rights to produce and sell a product for a certain period. Patents are intended to encourage innovation by allowing inventors to profit from their inventions. However, in a monopolized market, a patent holder has a significant advantage over potential competitors, as they cannot legally produce the patented product. This gives the patent holder a monopoly in the market.

Another form of legal restriction is licenses and other entry restrictions. These are legal requirements that must be met before a firm can enter a market. For example, in some countries, firms may be required to obtain a license before they can operate in a particular industry. These licenses may be difficult or expensive to obtain, making it challenging for new firms to enter the market.

For example, in India, the telecom industry is dominated by a few large players who hold the necessary licenses to operate. The high cost of acquiring these licenses acts as a significant barrier to entry for new firms, making it difficult for them to enter the market and compete with established players.

---

1. Monopoly



Another type of entry barrier is economies of scale. Economies of scale occur when a firm can supply market demand at a lower average cost per unit than two or more firms producing the same quantity of goods. This means that the market demand is not significant enough to allow more than one firm to achieve sufficient economies of scale. As a result, a single firm will emerge from the competitive process as the only supplier in the market.

There are two subtypes of economies of scale: natural monopoly and artificial monopoly. **A natural monopoly** occurs when a single firm can produce the entire market output at a lower average cost than any combination of two or more firms producing the same quantity of goods. This happens when there are high fixed costs in the industry, and the firm's average cost decreases as it produces more. A classic example of a natural monopoly is the electricity industry, where a single firm can generate and distribute electricity at a lower cost than multiple firms.

On the other hand, an **artificial monopoly** occurs when a single firm becomes dominant due to factors other than economies of scale. For example, a firm may become dominant by buying out or merging with its competitors or by engaging in anti-competitive practices. The Indian cement industry provides an example of an artificial monopoly, with a few large firms dominating the market through acquisitions and mergers.

---

1. Monopoly

The third type of entry barrier is the monopolist's control of an essential resource. This occurs when a firm has exclusive control over a critical input or resource required to produce a particular product. The firm can use this control to prevent competitors from entering the market, effectively establishing a monopoly.

For example, in India, the De Beers diamond company controlled a significant portion of the world's diamond supply, giving it significant power over the Indian diamond industry. The company's control over the diamond supply allowed it to manipulate the market and keep prices artificially high, effectively creating a monopoly.

---

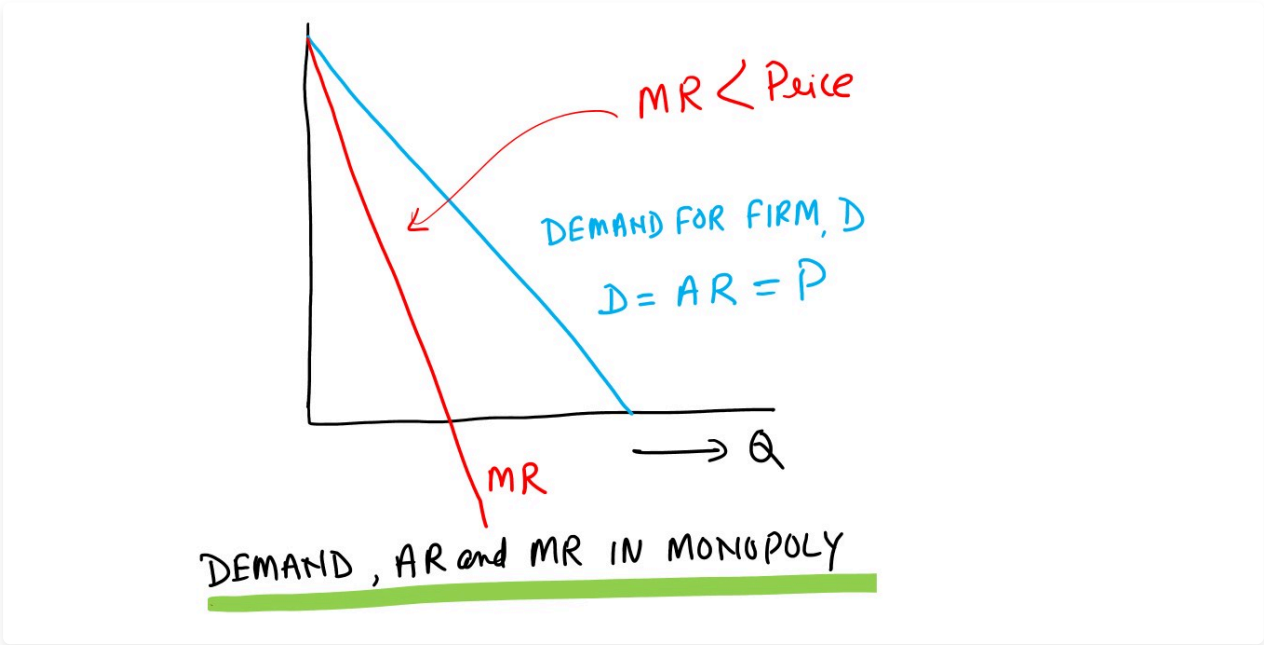
2. Monopsony

A **monopsony** market is one where there is only one buyer, who dominates the market. For example, the government is the single buyer of defense equipment in India. In a monopsony, the single buyer cannot purchase inputs at their desired price, as the price is determined by the market supply curve for the inputs. This is different from a monopoly market, where there is a single seller who can set the price for their product.

A **bilateral monopoly** is a combination of a monopoly and a monopsony, where there is a single buyer and seller in the market, such as in the labor market where the labor union is the monopoly seller negotiating with the employer as the monopsony buyer. In a bilateral monopoly, both the buyer and seller cannot maximize their profits simultaneously, and the price is determined by their relative bargaining power.

---

3. Demand, Price, Average Revenue and Marginal Revenue



Because a monopoly, by definition, supplies the entire market, the demand for goods or services produced by a monopolist is also the market demand. The demand curve for the monopolist's output therefore slopes downward, reflecting the law of demand—price and quantity demanded are inversely related.

Let's consider a hypothetical scenario where a company is the only provider of a certain medical treatment and has complete control over the market. The company has a demand curve that is downward-sloping, indicating that it can sell more treatments at lower prices and fewer treatments at higher prices.

Suppose the company can sell 5 treatments a day at a price of Rs. 10,000 each. The total revenue (TR) from selling 5 treatments is Rs. 50,000, which means that the average revenue (AR) per treatment is also Rs. 10,000. Therefore, the monopolist's price (P) is equal to its average revenue (AR) per unit.

If the company wants to sell the 6^{th} treatment, it must lower the price to Rs. 9,500. As a result, the TR for 6 treatments becomes Rs. 57,000, and the AR per treatment becomes Rs. 9,500.

The marginal revenue (MR) for the company from selling the 6^{th} treatment is Rs. 7,000 (Rs. 57,000 - Rs. 50,000). This value is less than the price (P) or average revenue (AR) of Rs. 9,500 per treatment.

To conclude, a monopolist's marginal revenue (MR) is always less than its price (P) or average revenue (AR), and it decreases as the quantity sold increases. (slope of MR is twice of slope of AR curve) This is because a monopolist must lower the price of all units sold to sell more units, resulting in a decrease in revenue from the previously sold units. This is shown in the figure.

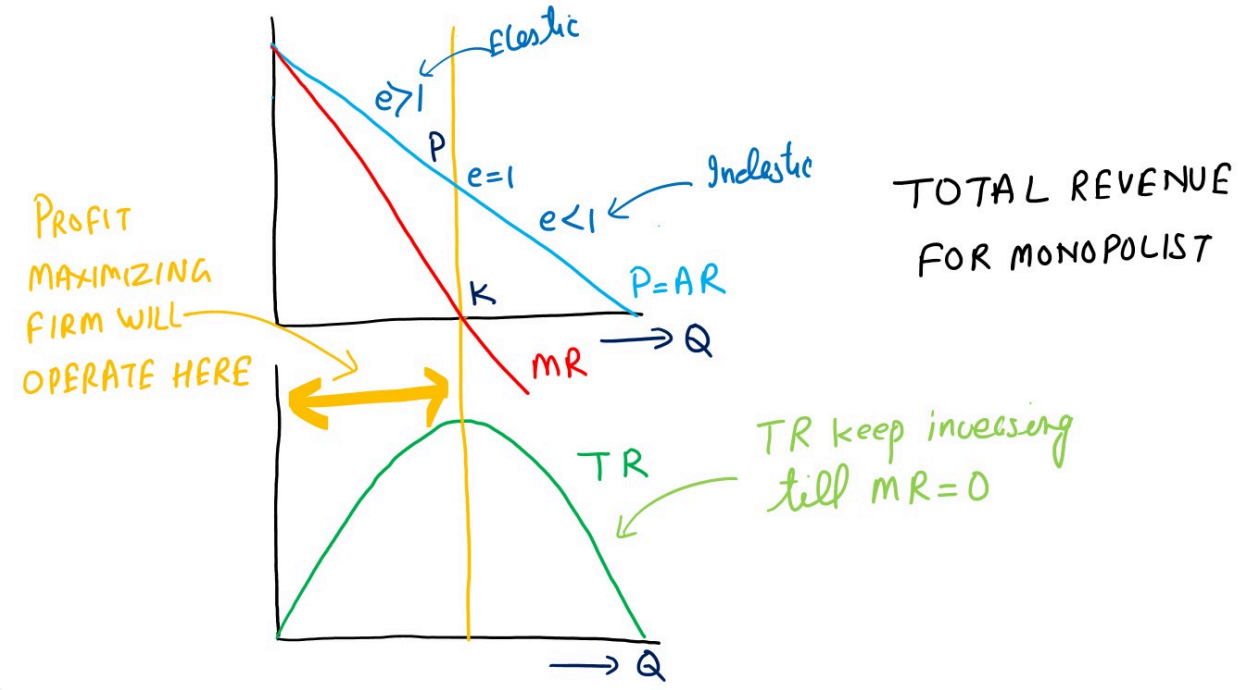
To sell more, the monopolist must lower the price on all units sold. Because the revenue lost from selling all units at a lower price must be subtracted from the revenue gained by selling another unit, MR is less than the P. At some point, MR turns negative.

As the price declines, the gap between P and MR widens because the loss from selling all products for less increases (because quantity increases) and the gain from selling another treatment decreases (because the price falls).

---

4. Total Revenue for Monopolist

The total revenue (TR) is nothing but equal to price (P) times quantity (Q). The TR reaches a maximum when MR reaches zero.



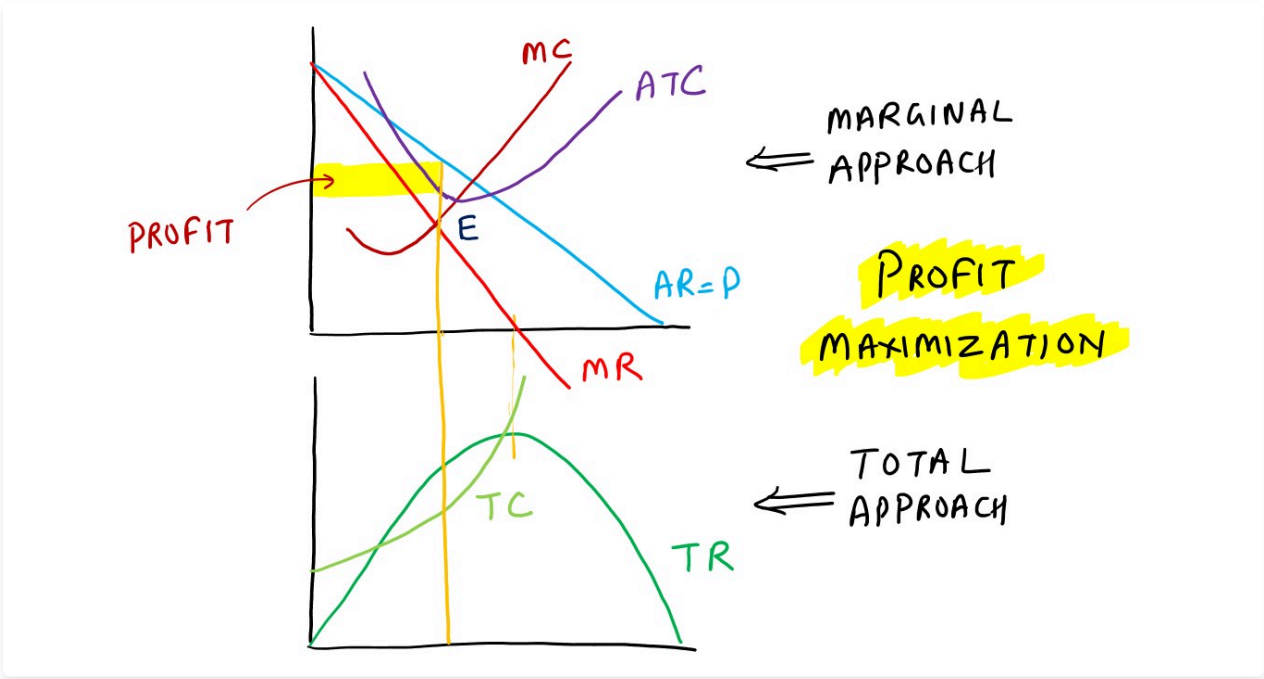
We have already learnt that the price elasticity for a straight-line demand curve decreases as you move down the curve. When demand is elastic—that is, when the percentage increase in quantity demanded more than offsets the percentage decrease in price—a decrease in price increases TR. Therefore, where demand is elastic, MR is positive, and TR increases as the price falls.

On the other hand, where demand is inelastic—that is, where the percentage increase in quantity demanded is less than the percentage decrease in price—a decrease in price reduces TR. In other words, the loss in revenue from selling all products for the lower price overwhelms the gain in revenue from selling more products. Therefore, where demand is inelastic, MR is negative, and TR decreases as the price falls.

As we can see in the figure, the MR turns negative if the price drops below point P, indicating inelastic demand below that price. A profit-maximizing monopolist would never willingly expand output to where demand is inelastic because doing so would reduce TR. It would make no sense to sell more just to see total revenue drop. Also note that demand is unit elastic at the price on the point P. At that price, MR is zero and TR reaches a maximum.

---

5. Profit Maximization



In the case of perfect competition, each firm's choice is confined to quantity because the market already determines the price. The perfect competitor is a **price taker**. The monopolist, however, can choose either the price or the quantity, but choosing one determines the other—they come in pairs. Because the monopolist can select the price that maximizes profit, we say the monopolist is a price maker.

One common myth about monopolies is that they charge the highest price possible. But the monopolist is interested in maximizing profit, not price. The monopolist's price is limited by consumer demand. So charging the highest possible price may not be consistent with maximizing profit.

As was the case with perfect competition, the monopolist can approach profit maximization in 2 ways:

1. **Total approach:** The monopolist produces the quantity at which total revenue (TR) exceeds total cost (TC) by the greatest amount.
2. **Marginal approach:** The profit-maximizing output occurs where marginal revenue (MR) equals marginal cost (MC), which is the golden rule of profit maximization.

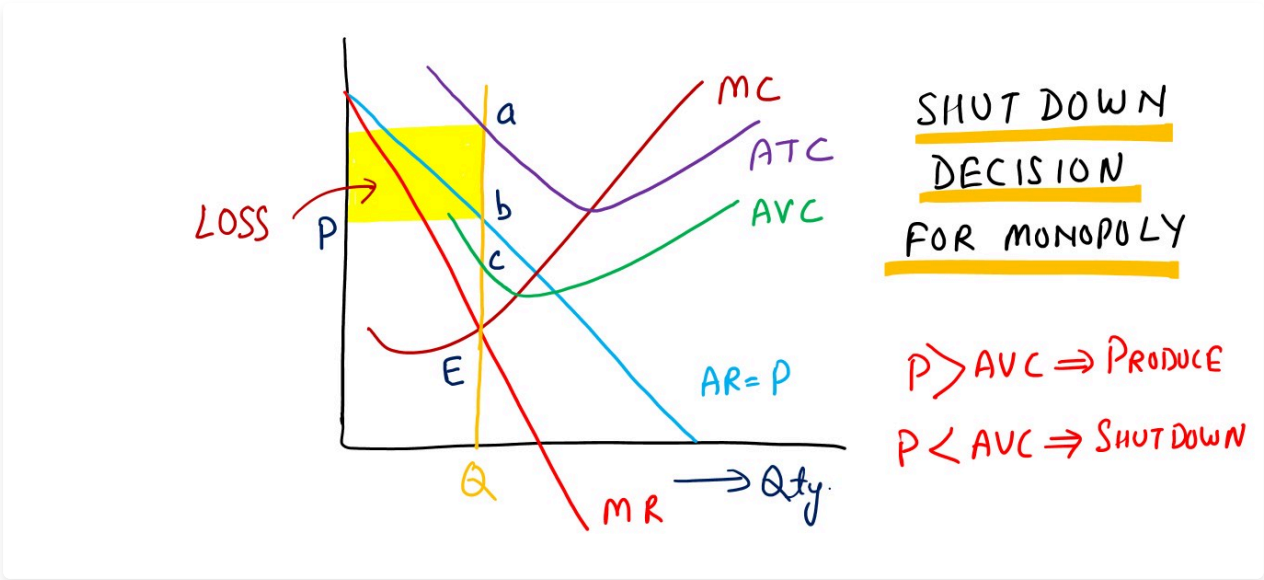
The cost and revenue data is shown in the figure. The intersection of the two marginal curves gives point E, indicating that profit is maximized at that quantity. So the profit-maximizing rate of output is found where the rising marginal cost (MC) curve intersects the marginal revenue (MR) curve.

The firm's profit or loss is measured by the vertical distance between the total revenue (TR) and total cost (TC) curves. The profit-maximizing firm will produce where TR exceeds TC by the greatest amount. Total profit can also be measured by the shaded area formed by multiplying average profit per unit by the number of units sold (in case of marginal analysis).

---

6. Short-Run Losses and the Shutdown Decision

A monopolist is not assured a profit. Although a monopolist is the sole supplier of a good with no close substitutes, the demand for that good may not generate economic profit in either the short run or the long run.



In the short run, the loss-minimizing monopolist, like the loss-minimizing perfect competitor, must decide whether to produce or to shut down. If the price (P) covers average variable cost (AVC), the firm will produce. If not, the firm will shut down, at least in the short run.

We know that average variable cost (AVC) and average fixed cost (AFC) sum to average total cost (ATC). Loss minimization occurs, where the marginal revenue (MR) curve intersects the marginal cost (MC) curve. This is shown by point E in the figure.

At the equilibrium rate of output, Q, Price P is found on the demand curve at point b. That Price (P) exceeds average variable cost (AVC), at point c, but is below average total cost (ATC), at point a. Because price (P) covers average variable cost (AVC) and makes some contribution to average fixed cost, this monopolist loses less by producing Q than by shutting down. The average loss per unit, measured by ab, is average total cost minus average revenue, or price. The loss, identified by the shaded rectangle, is the average loss per unit, ab, times the quantity sold, Q. The firm will shut down in the short run if the average variable cost curve is above the demand curve, or average revenue curve, at all output rates.

Recall that a perfectly competitive firm's supply curve is that portion of the marginal cost curve at or above the average variable cost curve. The intersection of a monopolist's marginal revenue and marginal cost curves identifies the profit-maximizing (or loss-minimizing) quantity, but the price is found up on the demand curve. Because the equilibrium quantity can be found along a monopolist's marginal cost curve, but the equilibrium price appears on the demand curve, no single curve shows both price and quantity supplied. Because no curve reflects combinations of price and quantity supplied, there is no monopolist supply curve.

---

7. Long run Profit Maximization

For perfectly competitive firms, the distinction between the short run and the long run is important because entry and exit of firms can occur in the long run, erasing any economic profit or loss. For the monopolist, the distinction between the short run and long run is less important. If a monopoly is insulated from competition by high barriers that block new entry, economic profit can persist in the long run. Yet short-run profit is no guarantee of long-run profit.

---

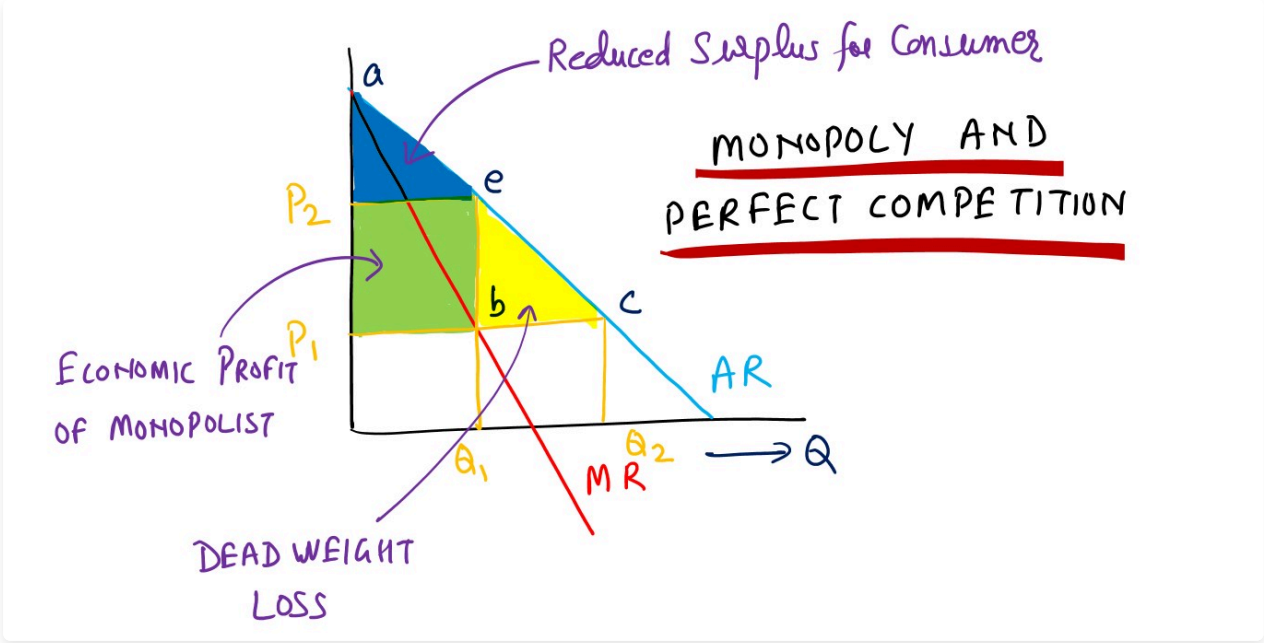
8. Comparing Monopoly and Perfect Competition

It may be noted now that the monopolists are no greedier than perfect competitors (because both maximize profit). Also, the monopolists do not charge the highest possible price. Further, even monopolists are not guaranteed a profit.

A perfectly competitive industry would produce output Q_2, determined by the intersection of the market demand curve and the market supply curve. The price would be P_1. A monopoly that could produce output at the same minimum average cost as a perfectly competitive industry would produce output Q_1, determined by point, where marginal cost (MC) and marginal revenue (MR) intersect. The monopolist would charge price P_2.

Thus, given the same costs, output is lower and price is higher under monopoly than under perfect competition.

Let us consider the allocative and distributive effects of monopoly versus perfect competition. In the figure, consumer surplus under perfect competition is the large triangle, acP_1. Under monopoly, consumer surplus shrinks to the smaller triangle aeP_2. The monopolist earns economic profit equal to the shaded rectangle.



By comparing the situation under monopoly with that under perfect competition, you can see that the monopolist's economic profit comes entirely from what was consumer surplus under perfect competition. Because the profit rectangle reflects a transfer from consumer surplus to monopoly profit, this amount is not lost to society and so is not considered a welfare loss.

Notice, however, that consumer surplus has been reduced by more than the profit rectangle. Consumers have also lost the triangle ebc, which was part of the consumer surplus under perfect competition. The ebc triangle is called the **deadweight loss** of monopoly because it is a loss to consumers but a gain to nobody. The deadweight loss is the net loss to society when a firm uses its market power to restrict output and increase price. This loss results from the allocative inefficiency arising from the higher price and reduced output of monopoly. Again, society would be better off if output exceeded the monopolist's profit-maximizing quantity, because the marginal benefit of more output exceeds its marginal cost. Under monopoly, the price, or marginal benefit, always exceeds marginal cost.

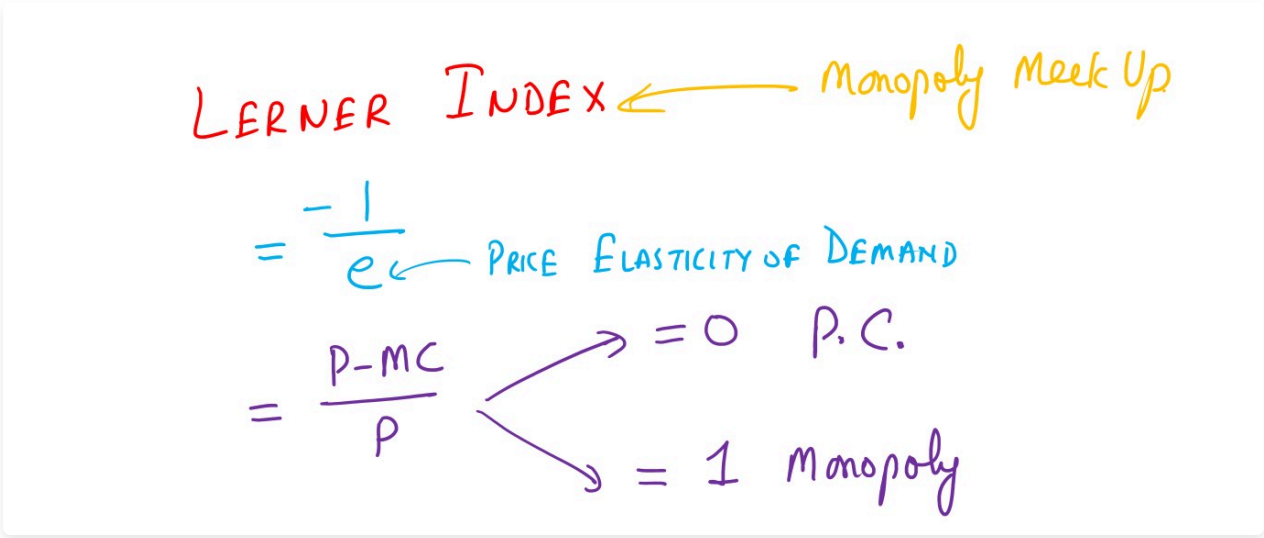
---

9. Measures of Monopoly

Let us discuss a few measures, which are used to gauage the level of competition in the market.

---

9. Measures of Monopoly



The Lerner index is a measure of monopoly power for it shows the effect of price elasticity of demand on monopoly price, relative to the product marginal cost. It is named after Abba Lerner and defined as the ratio of the difference between price and marginal cost relative to the price. It can also be considered the negative inverse of the price elasticity of demand.

Lerner Index =$\frac{P-MC}{P}=-\frac{1}{e}$

It is also called **Elasticity reciprocal**.

The Lerner index takes a value between 0 and 1, but greater the index, greater the difference between price and marginal cost in favor of the price, which would mean a greater monopoly power. This is naturally confirmed in the case of the perfect competition where the difference between price and marginal cost is zero (P = MR = MC), which explains that in a perfect competition there is no monopoly power at all.

In competitive markets, P = MC, so the markup on price tends to converge toward zero as competitive pressures increase. Conversely, P > MC in monopoly markets, so the markup on price can be expected to rise as competitive pressures decrease.

---

9. Measures of Monopoly

Concentration ratios are used as measures of monopoly power in industries. These ratios provide insights into the extent to which sales, profits, or assets are concentrated among a few large firms versus smaller ones within a particular industry. Concentration ratios help assess the level of market competition and whether monopoly power exists.

A *concentration ratio* is a measure of the market share held by the largest firms in an industry relative to the total market. It is expressed as a percentage. The concentration ratio is calculated by adding the market shares of the largest firms in descending order until a certain number of firms (denoted as "n") are considered. The resulting ratio, known as CR_n, represents the cumulative market share of the top "n" firms.

If there are five firms in an industry, the CR_2 represents the cumulative market share of the two largest firms, CR_3 represents the cumulative market share of the top three firms, and so on.

Higher concentration ratios indicate that a smaller number of firms dominate the industry, suggesting a greater degree of market power. Conversely, lower concentration ratios imply a more competitive market with sales, profits, or assets distributed more evenly among firms.

Concentration ratios do not consider the total number of firms in an industry. Thus, industries with similar concentration ratios may have different competitive dynamics due to varying numbers of firms.

---

9. Measures of Monopoly

The Herfindahl Index (HI) given by Herfindahl and Hirschmann (also called Herfindahl–Hirschmann index (HHI)), is a measure used to assess market concentration within an industry. Unlike concentration ratios (CRs), the Herfindahl Index takes into account both the number of firms in the industry and their relative market shares. It provides a more comprehensive picture of market structure and is particularly useful for analyzing market power and competition.

HI =\(\sum_1^n{S_i^2} \)

where, where n is the number of firms in the industry and is the market share of the ith firm (i = 1, 2, ..., n). If there are 5 firms with market shares of 50%, 30%, 10%, 6% and 4%, then HI will be:

$HI = 0.50^2 + 0.30^2 + 0.10^2 + 0.06^2 + 0.04^2 = 0.3552$

In case all the firms had equal market shares of 20%, the Herfindahl Index would be

$HI = 5\ (0.20^2) = 1/5$

That is, if there are n firms in an industry all having equal shares, the share of each firm would be 1/n. In the case of pure monopoly, the HI would be equal to 1, and it is the maximum value of HI.

Thus, the HI would lie between 1/n and 1, both ends inclusive ($1/n \leq HI \leq 1$), and a larger HI indicates a greater monopoly power.

If value of S_i are taken in percentage, then maximum value of HI will be 10,000.

---

1. Price Discrimination

Till now, we have understood that, a monopolist, to sell more output, must lower the price. In reality, a monopolist can sometimes increase profit by charging higher prices to those who value the product more. This practice of charging different prices to different groups of consumers is called **price discrimination**.

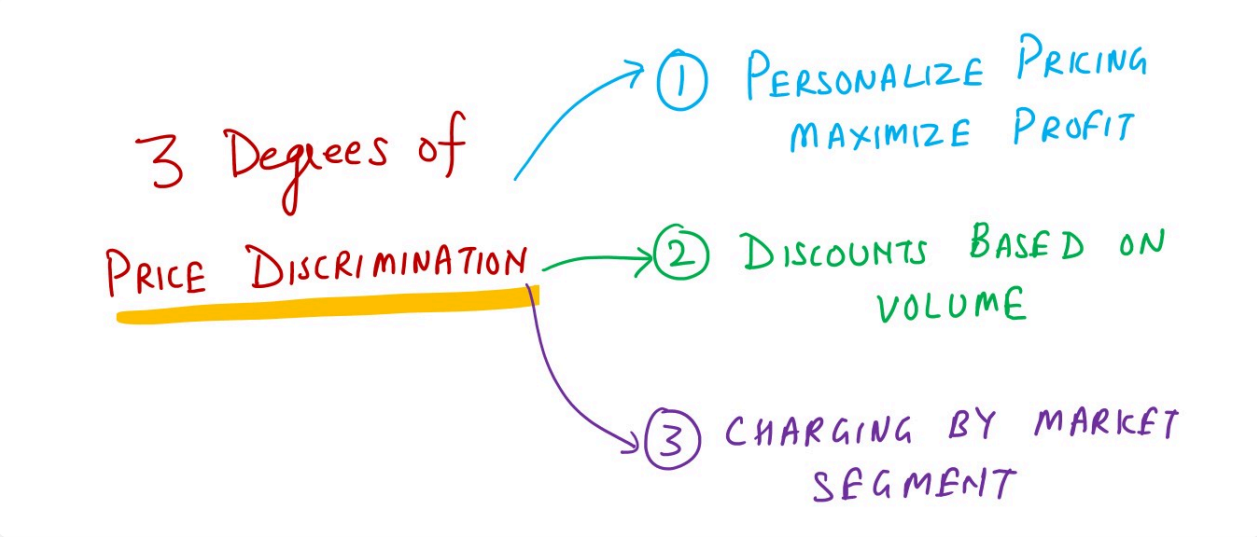
An airline company offers lower prices for tickets booked in advance, compared to tickets booked closer to the travel date. This is an example of price discrimination based on the assumption that some customers are willing to plan ahead and make early bookings to save money, while others are willing to pay more for last-minute bookings due to urgency or convenience.

To practice price discrimination, a firm's product must meet certain conditions:

1. The demand curve for the firm's product must slope downward, indicating that the firm is a price maker—the producer has some market power, some control over the price. There should be imperfect competition.
2. There must be at least two groups of consumers for the product, each with a different price elasticity of demand.
3. The firm must be able, at little cost, to charge each group a different price for essentially the same product.
4. The firm must be able to prevent those who pay the lower price from reselling the product to those who pay the higher price.

2. Degrees of Price Discrimination

There are 3 degrees of price discrimination.



Such a pricing scheme that makes each consumer pay the maximum amount that he is willing to pay is known as **first degree price discrimination.** Under this method of pricing, firm separates the market into each individual consumer & charge price which they are able and willing to pay. Auctions are the best example of first degree price discrimination. They enable the seller to identify those consumers who are willing to pay the maximum possible price for their product. This is also called **perfect price description**. The attempt is to minimize Consumer Surplus.

Where auctions are not feasible, firms try their best to approximate the first degree price discrimination. There are two ways of doing this. The first approach is based on the law of diminishing marginal utility. Accordingly, it believes that the satisfaction that an individual consumer derives from each successive unit of a commodity goes on diminishing as he consumes more of it. The incremental value that a consumer perceives in a product gets lower with each additional unit consumed. A thirsty Coke-loving consumer will pay the maximum amount possible for the first can of Coke. His willingness to pay for all additional units will gradually decrease. The amount that he will be willing to pay for the fifth can, will not be as much as he did for the first can. All the assumptions of the law of diminishing marginal utility hold good in this case.

Even the price paid for the first unit will be different for different consumers depending upon their love for Coke. The price at which one consumer buys Coke may not attract other consumers. So the seller comes up with a scheme wherein he offers Coke at a lower unit price in packs of six cans each. The consumer gets the benefit of reduced prices only if he buys the full pack with six cans. Thus, the consumer who needs Coke badly but needs just one unit, can have it by paying a higher price, while the other consumer who is not that much desperate can get a lower price by buying a pack of six cans. Such a form of price discrimination, which is based on the **volume of consumer purchase**, is known as **second degree price discrimination.** It is quite commonly found in everyday life. Quantity discounts on products is a classic illustration of this pricing practice.

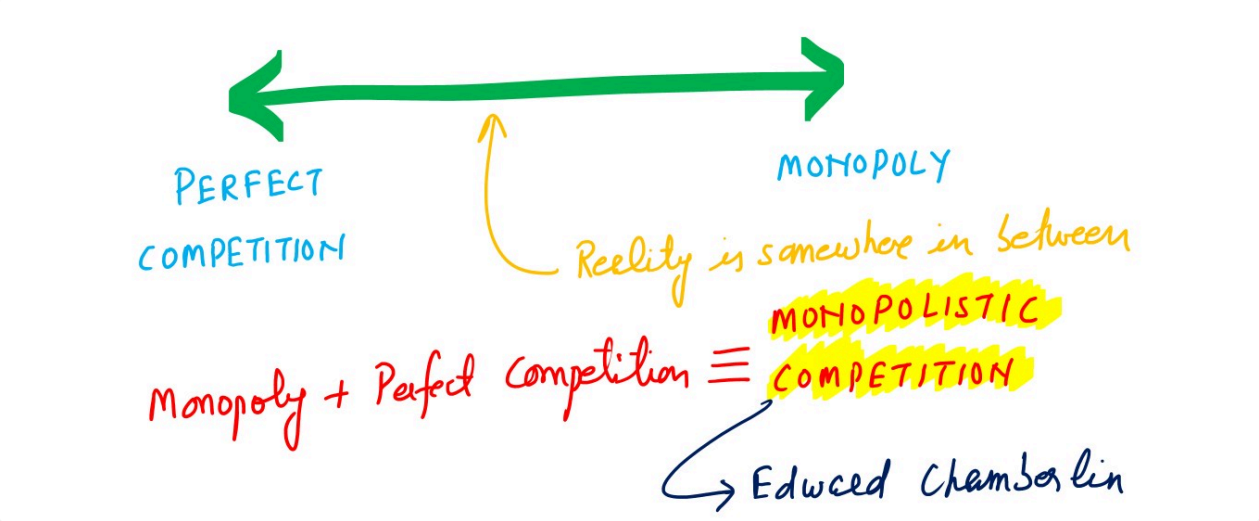
The second way of approximating first degree price discrimination is where products are priced according to the **type of buyer** and not the volume of purchases. Such a pricing mechanism where pricing is based on the characteristics of buyer is known as **third degree price discrimination**.

This type of price discrimination occurs when an imperfectly competitive firm distinctively segments the market into different segments and charges them different prices for the same product. Classic examples of this price discrimination are the seemingly goodwill-driven and community-spirited discounts that are offered to seniors, kids, students, veterans, or any other market segment that is distinguished by a certain characteristic such as age, profession, marital status, and alike. This can also be clear in the market segments distinguished by time of consumption, especially in the case of movie tickets or vacation packages. It is assumed that charging different prices for different markets is practiced until the marginal revenue in each market equals the marginal cost of producing the product.

---

1. Introduction

During the 1920s and 1930s, economists developed models that fell between perfect competition and monopoly, resulting in two models of monopolistic competition. In 1933, Edward Chamberlin published *The Theory of Monopolistic Competition*, while Joan Robinson published *The Economics of Imperfect Competition* the same year.

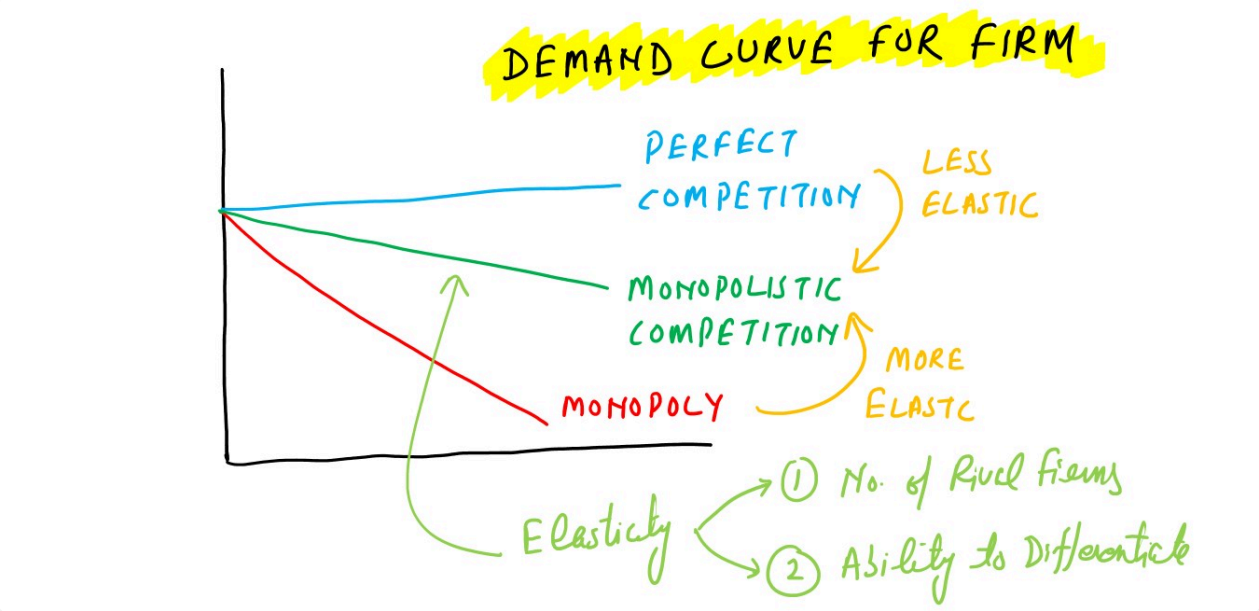


Monopolistic competition is a market structure in which many firms sell products that are substitutes, but different enough that each firm's demand curve slopes downward, and entry into the market is relatively easy. This market structure combines elements of both monopoly and competition. Chamberlin used the term *monopolistic competition* to describe a market in which many producers offer products that are substitutes but are not viewed as identical by consumers.

---

2. Demand Curve for Firm

In monopolistic competition, each supplier has some power over the price it can charge because the products of different suppliers differ slightly. Therefore, the demand curve for each supplier slopes downward. As a result, firms in monopolistic competition are not price takers but are price makers.

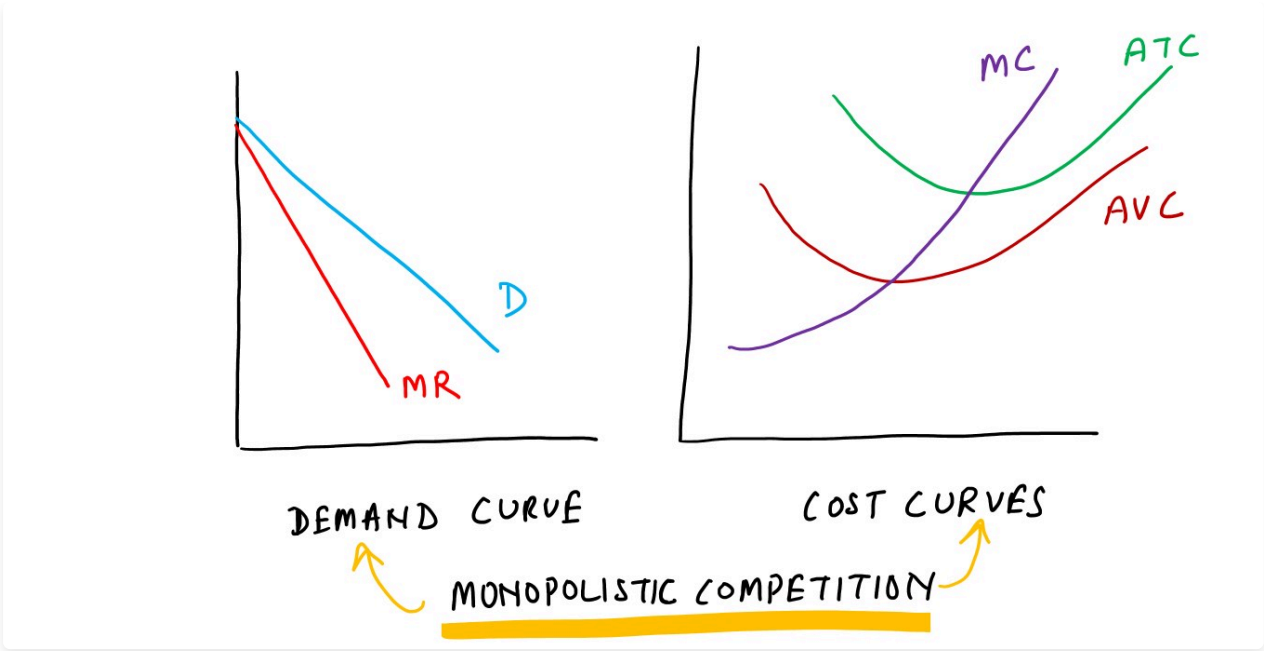


Moreover, because barriers to entry are low, firms in monopolistic competition can enter or leave the market with ease. Thus, there are enough sellers in this market that they behave competitively, but there are also enough sellers that each tends to get lost in the crowd.

Because many firms are selling substitutes, any firm that raises its price can expect to lose some customers, but not all, to rivals. By way of comparison, a price hike would cost a monopolist fewer customers but would cost a perfect competitor all customers. Therefore, a monopolistic competitor faces a demand curve that tends to be more elastic than a monopolist's but less elastic than a perfect competitor's.

We know that the availability of substitutes for a given product affects its price elasticity of demand. The price elasticity of the monopolistic competitor's demand depends on:

1. the number of rival firms that produce similar products and
2. the firm's ability to differentiate its product from those of its rivals.



A firm's demand curve will be more elastic the more substitutes there are and the less differentiated its product is.

---

3. Differentiation

Sellers differentiate their products in 4 basic ways:

1. **Physical Differences:** The most obvious way products differ is in their physical appearance and qualities, such as size, weight, color, taste, texture, and packaging design. Shampoos, for example, differ in color, scent, thickness, lathering ability, and bottle design.
2. **Location:** The number and variety of locations where a product is available are other ways of differentiation, such as spatial differentiation. For example, convenience stores are likely to be nearer customers, have no long lines, and are open all night.
3. **Services:** Products also differ in terms of their accompanying services, such as product demonstrations, online support, toll-free numbers, and money-back guarantees.
4. **Product Image:** A final way products differ is in the image the producer tries to foster in the consumer's mind. For example, some producers try to demonstrate high quality based on where products are sold or tout their all-natural ingredients.

   Imagine you're in a market with multiple pizza outlets. Each pizza place has its own specialty toppings, crust styles, and delivery options. This is an example of monopolistic competition. Each pizza place differentiates itself from its competitors by offering unique pizza options and services, like gluten-free crust, extra cheese, vegan toppings, 30 minutes home delivery or 24/7 delivery. Since there are many pizza places in the city, they all compete with each other, but each can charge slightly different prices for their products and services because they are not perfect substitutes for one another.

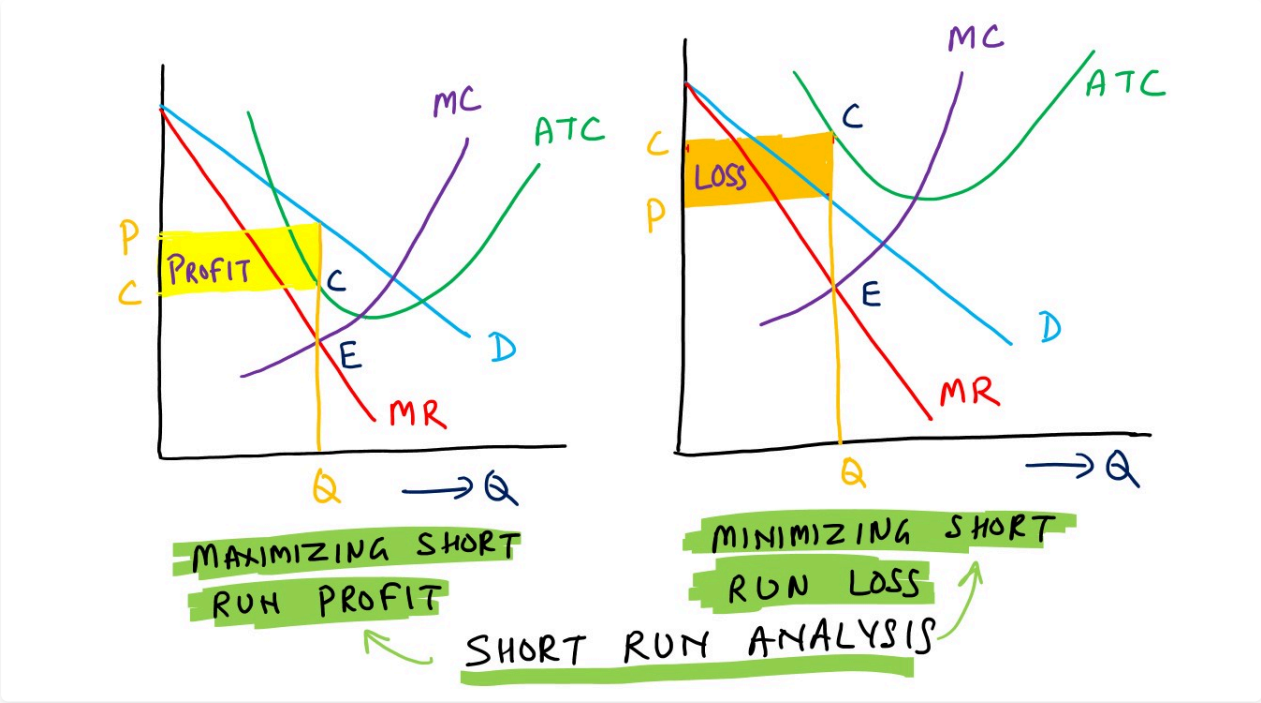
---

4. Short-Run Profit Maximization or Loss Minimization

In the case of monopolistic competition, just like monopoly, the downward-sloping demand curve (D) means the marginal revenue (MR) curve also slopes downward and lies beneath the demand curve. The figure depicts D and MR curves for a monopolistic competitor. The figure also presents average and marginal cost curves. Remember that the forces that determine the cost of production are largely independent of the forces that shape demand, so there is nothing special about a monopolistic competitor's cost curves.

In the short run, a firm that can at least cover its variable cost will increase output as long as marginal revenue (MR) exceeds marginal cost (MC). A monopolistic competitor maximizes profit just as a monopolist does: the profit-maximizing quantity occurs where marginal revenue equals marginal cost; the profit maximizing price for that quantity is found up on the demand curve.

The figure shows that the marginal cost (MC) and marginal revenue (MR) curves intersect at point E, yielding equilibrium output Q, equilibrium price P.



As shown in the figure, at the firm's profit-maximizing quantity (Q), average total cost (C), is below the price (P). Price minus average total cost is the firm's profit per unit (P minus C), which, when multiplied by the quantity (Q), yields economic profit, shown by the shaded rectangle. Again, the profit-maximizing quantity is found where MR equals MC; price is found up on the demand curve at that quantity. Thus, a monopolistic competitor, like a monopolist, has no supply curve—that is, there is no curve that uniquely relates alternative prices and corresponding quantities supplied.

The monopolistic competitor, like other firms, has no guarantee of economic profit. If the average total cost (ATC) curve lies entirely above the demand (D) curve, no quantity would allow the firm to break even. Thus, the firm suffers a short-run loss equal to the loss per unit (C minus P) multiplied by Q, shown by the shaded rectangle. In such a situation, the firm must decide whether to produce or to shut down temporarily. The rule here is the same as with perfect competition and monopoly: as long as the price (P) exceeds average variable cost (AVC), the firm in the short run will lose less by producing than by shutting down. If no price covers average variable cost, the firm will shut down.

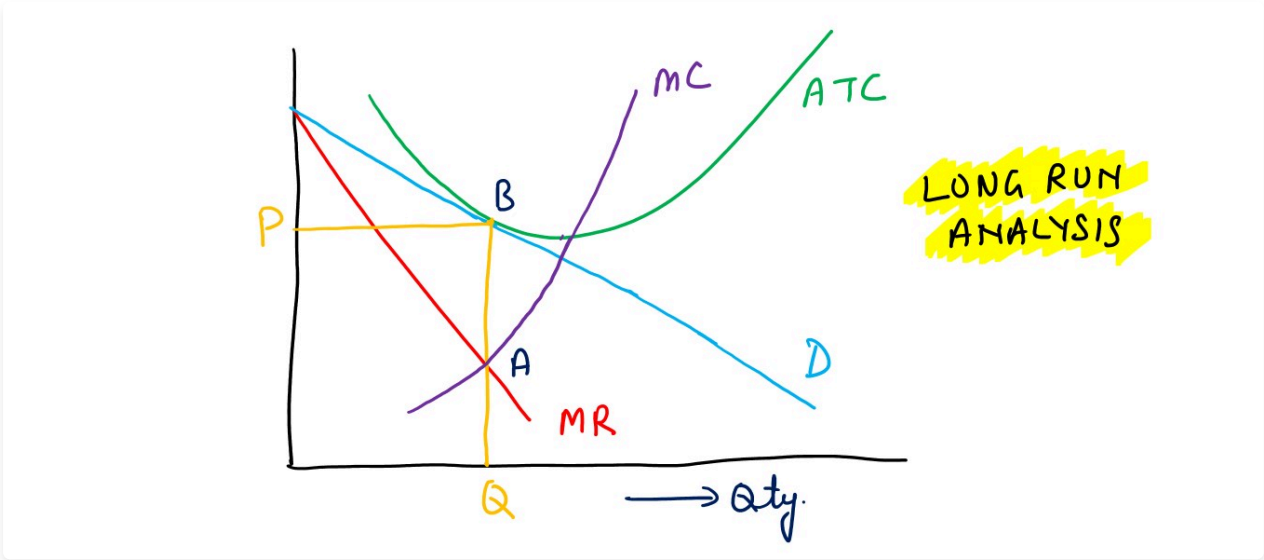
Thus, the short-run profit maximization (or loss minimization) in monopolistic competition is quite similar to that under monopoly.

---

5. Long Run analysis

Low barriers to entry in monopolistic competition mean that short-run economic profit will attract new entrants in the long run. Because new entrants offer products that are similar to those offered by existing firms, they draw customers away from existing firms, thereby reducing the demand facing each firm. Entry will continue in the long run until economic profit disappears. Because of the ease of entry to the market, monopolistically competitive firms earn zero economic profit in the long run.

Similarly, if they suffer short-run losses, some monopolistic competitors will leave the industry in the long run, redirecting their resources to products expected to earn at least a normal profit. As firms leave, their customers will switch to the remaining firms, increasing the demand for those products. Firms will continue to leave in the long run until the remaining firms have sufficient customers to earn normal profit, but not economic profit.



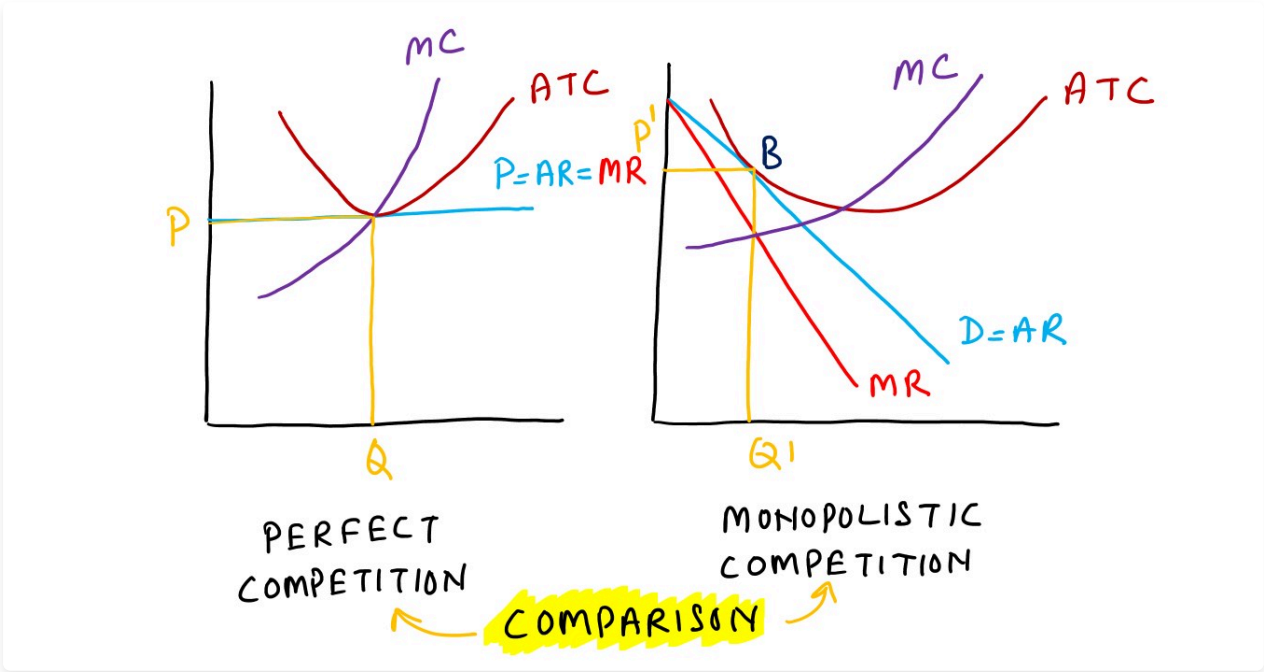
The figure shows long-run equilibrium for a typical monopolistic competitor. In the long run, entry and exit will alter each firm's demand curve until economic profit disappears—that is, until price (P) equals average total cost (ATC). The marginal revenue (MR) curve intersects the marginal cost (MC) curve at point A. At the equilibrium quantity (Q), the average total cost (ATC) curve at point B is tangent to the demand (D) curve. Because average total cost (ATC) equals the price (P), the firm earns no economic profit but does earn a normal profit. At all other rates of output, the firm's average total cost (ATC) is above the demand (D) curve, so the firm would lose money if it reduced or expanded its output.

To conclude, the monopolistic competition is like monopoly in the sense that firms in each industry face demand curves that slope downward. Monopolistic competition is like perfect competition in the sense that easy entry and exit eliminate economic profit or economic loss in the long run.

---

6. Comparing Perfect Competition Versus Monopolistic Competition in Long-Run

Let us compare monopolistic competition with perfect competition in terms of efficiency. In the long run, neither can earn economic profit. The difference arises because of the different demand curves facing individual firms in each of the two market structures. The figure shows the long-run equilibrium price and quantity for a typical firm in each of two market structures, assuming each firm has identical cost curves. In each case, the marginal cost (MC) curve intersects the marginal revenue (MR) curve at the quantity where the average total cost (ATC) curve is tangent to the firm's demand (D) curve.



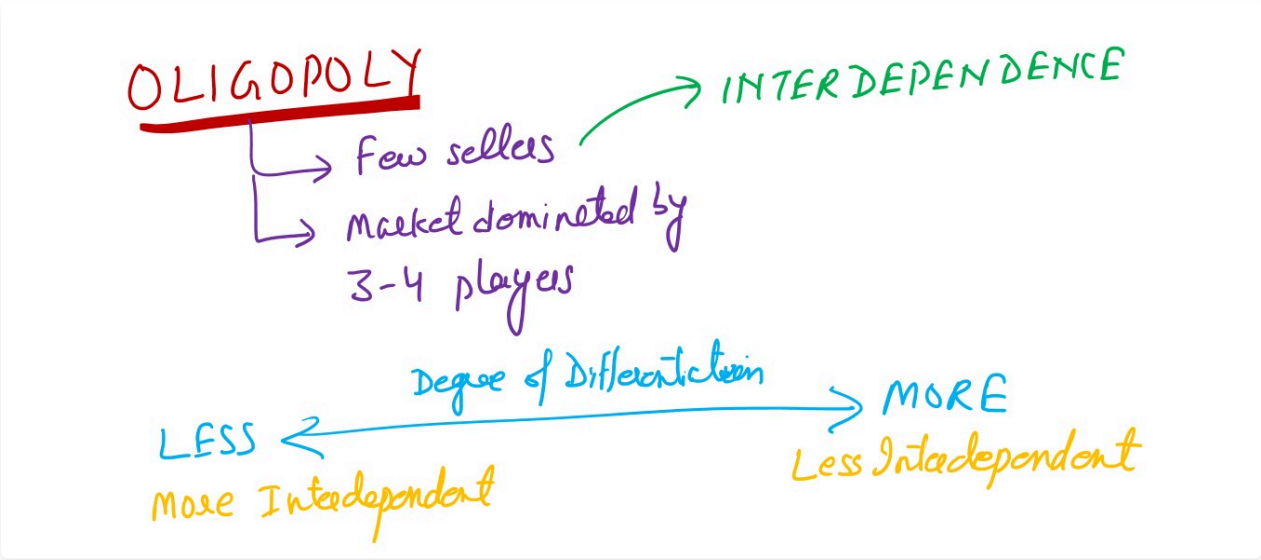
The perfectly competitive firm faces a demand (D) curve that is horizontal at market price P. Long-run equilibrium occurs at output Q, where the demand (D) curve is tangent to the average total cost curve (ATC) at its lowest point. The monopolistically competitive firm is in long run equilibrium at output Q', where demand (D) is tangent to average total cost (ATC). Because the demand curve slopes downward in case of monopolistically competitive firm, however, the tangency does not occur at the minimum point of average total cost (ATC). Thus, the monopolistically competitive firm produces less output and charges a higher price than does a perfectly competitive firm with the same cost curves. Neither firm earns economic profit in the long run.

Firms in monopolistic competition are not producing at minimum average cost. They are said to have **excess capacity**, because production falls short of the quantity that would achieve the lowest average cost. Excess capacity means that each producer could easily serve more customers and in the process would lower average cost. The marginal value of increased output would exceed its marginal cost, so greater output would increase social welfare.

One other difference between perfect competition and monopolistic competition does not show up in the figure. Although the cost curves drawn are identical, firms in monopolistic competition advertise more to differentiate their products than do firms in perfect competition. These higher advertising costs shift up their average cost curves.

1. Oligopoly

Oligopoly is a market structure in which only a few large firms dominate the market, and these firms are interdependent and must consider the impact of their actions on their competitors. In India, several industries operate in an oligopolistic market, such as the automobile industry, where a few large players such as Maruti Suzuki, Hyundai, and Tata Motors dominate the market share. Similarly, the Indian steel industry is also an oligopolistic market with three main players - Tata Steel, JSW Steel, and SAIL - accounting for a significant share of the market.



Due to the few firms in an oligopoly (Greek word meaning ‘few sellers’), each firm's pricing and production decisions can have a significant impact on the other firms in the market. For example, if one automobile company lowers its prices, it may cause a price war with other firms that can ultimately result in lower profits for all companies involved.

Oligopolistic firms may also engage in non-price competition, such as advertising or product differentiation, to attract customers and gain a competitive edge. For example, automobile companies may advertise their cars' safety features or luxury amenities to differentiate themselves from their competitors.

Undifferentiated oligopoly is a market structure in which the firms sell a commodity, which means that the product does not differ across suppliers. For example, the steel industry in India is an undifferentiated oligopoly as steel produced by one company is not significantly different from the steel produced by other companies. In this type of oligopoly, firms compete mainly on price and quantity, as there are no other factors that can differentiate their products.

On the other hand, **differentiated oligopoly** is a market structure where firms sell products that differ across suppliers. For example, the automobile industry in India is a differentiated oligopoly as different car manufacturers produce cars that differ in terms of design, features, performance, and other attributes. In this type of oligopoly, firms compete based on product differentiation, marketing, and other factors, apart from price and quantity.

Because of **interdependence**, the behavior of any particular firm is difficult to predict. Each firm knows that any changes in its product’s quality, price, output, or advertising policy may prompt a reaction from its rivals. And each firm may react if another firm alters any of these features.

Duopoly refers to a market structure where there are only two sellers or firms that dominate the market. The concept of the prisoner's dilemma can be applied to duopolies in terms of pricing strategies.

Analyzing the behavior of oligopolists is a complex task due to their interdependence. There is no single model or approach that can completely explain their behavior. Oligopolists can either coordinate their actions to act as a single monopolist or compete so fiercely that it leads to price wars. Various theories have been developed to explain oligopoly behavior, including collusion, price leadership, and game theory. Each approach has some relevance in explaining observed behavior, but none is entirely satisfactory as a general theory of oligopoly. Instead, there is a set of theories based on the diversity of observed behavior in an interdependent market.

---

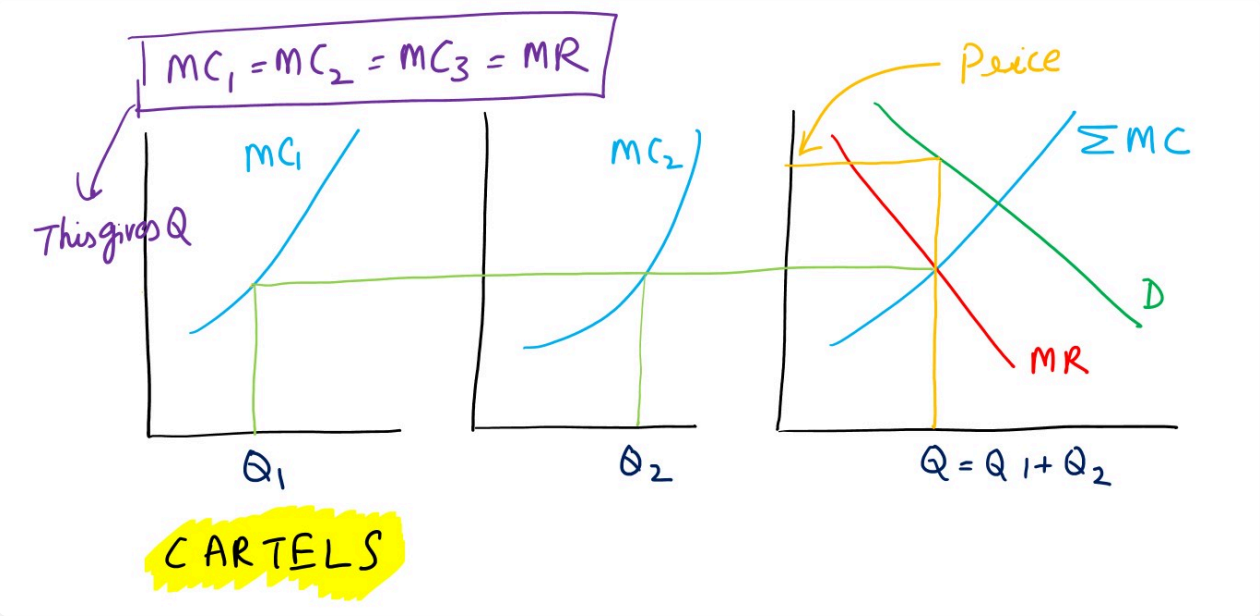
1. Oligopoly

The following are the barriers to entry that lead to the creation of oligopoly:

1. **Economies of Scale**: Large firms can produce goods at a lower cost per unit than small firms due to economies of scale. This can create a barrier to entry for smaller firms that cannot produce goods as efficiently as larger firms.
2. **Legal Restrictions**: Some industries are subject to legal restrictions that can make it difficult for new firms to enter the market. For example, the pharmaceutical industry is heavily regulated, and obtaining approval for a new drug can be a long and expensive process.
3. **Brand Names Built up by Years of Advertising**: Established firms with well-known brand names and a loyal customer base can make it difficult for new firms to enter the market. This is because consumers may prefer to buy from brands they know and trust.
4. **Control Over an Essential Resource**: Firms that control essential resources, such as raw materials or distribution channels, can create a barrier to entry for new firms that need access to these resources.
5. **Crowding Out the Competition**: In some cases, existing firms may engage in aggressive pricing strategies or other tactics to drive new firms out of the market. This can make it difficult for new firms to gain a foothold and compete with established firms.

2. Collusion and Cartels

In an oligopolistic market, there are just a few firms so, to decrease competition and increase profits, they may try to collude, or conspire to rig the market. **Collusion** is an agreement among firms in the industry to divide the market and fix the price. A **cartel** is a group of firms that agree to collude so they can act as a monopoly to increase economic profit. Cartels are more likely among sellers of a commodity, like oil or steel. Colluding firms, compared with competing firms, usually produce less, charge more price, block new firms, and earn more economic profit.



Collusion and Cartels are illegal in most of the countries.

A cartel acts like a monopolist. In the figure, D is the market demand curve, MR the associated marginal revenue curve, and MC the summation of the marginal cost curves of cartel members (assuming all firms in the market join the cartel).

Just like Monopoly, the cartel profits are maximized when the industry produces quantity Q (given by MR = MC) and charges corresponding price P on demand curve, D.

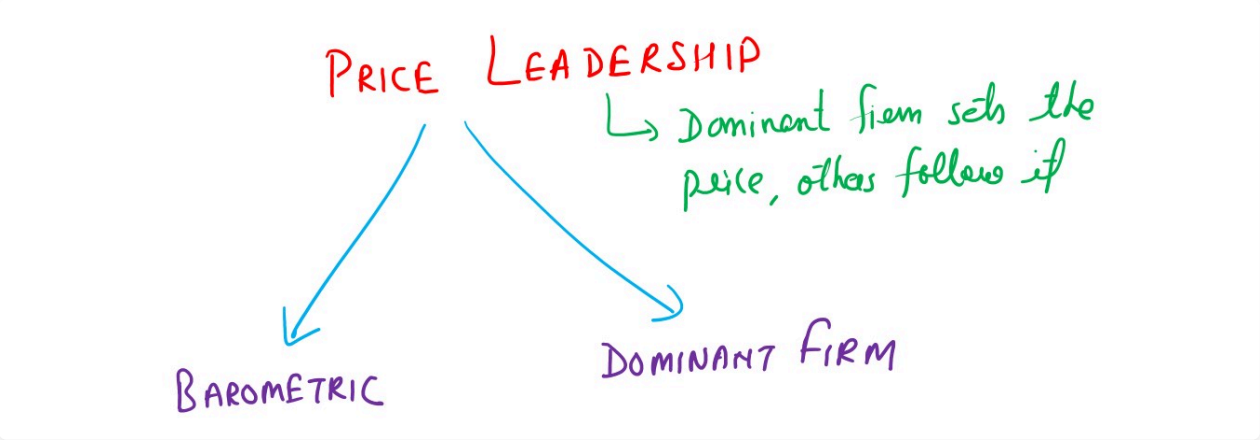
To maximize cartel profit, output Q must be allocated among cartel members so that each member's marginal cost equals C (C = MC_1 =MC_2 = MR). Any other allocation would lower cartel profit. Thus, for cartel profit to be maximized, output must be allocated so that the marginal cost for the final unit produced by each firm is identical.

However, there are various challenge:

1. If there are differences in the average costs among firms, the allocation that maximizes cartel profit will result in unequal profits for each firm, which may lead to some firms leaving the cartel.
2. As the number of firms in the cartel grows, reaching a consensus becomes more challenging.
3. New entry into the market can disrupt the cartel and decrease economic profit unless the cartel can prevent new firms from entering.
4. Cheating on the agreement is a common issue that undermines the cartel's efforts to maintain their arrangement.
5. The product differs among firms, making it difficult to reach an agreement on the allocation of output.

---

3. Price Leadership



Price leadership is a form of informal or tacit collusion that occurs in an oligopoly market where a dominant firm sets the price for the rest of the industry. Other firms then follow that lead, which helps them avoid price competition. This practice is common in industries with a few large players and significant barriers to entry. The price leader typically initiates any price changes, and others follow.

Price leadership is not without its obstacles. The practice is illegal in some countries. Secondly, the greater the product differentiation among sellers, the less effective price leadership will be as a means of collusion. Third, there is no guarantee that other firms will follow the leader, as firms that fail to follow a price increase take business away from firms that do. Fourth, unless there are barriers to entry, a profitable price will attract new entrants, which could destabilize the price-leadership agreement. Finally, as with formal cartels, some firms are tempted to cheat on the agreement to boost sales and profits.

In India, the cement industry is an example of price leadership. Cement companies in India have been known to engage in price leadership, where the largest firm in the industry sets the price for the rest of the industry. For example, in 2019, UltraTech Cement, the largest cement producer in India, announced a price increase of Rs 30-50 per bag. Other cement companies followed suit, raising prices in a similar range. However, ACC and Ambuja Cements, both subsidiaries of Swiss cement company Holcim, did not follow the price increase, leading to a price war within the industry. This example highlights the challenges of maintaining price leadership in an oligopoly market, even in the face of a dominant firm.

There are two primary patterns of price leadership seen in various industries: barometric and dominant price leadership.

In the **barometric price leadership**, a firm takes the initiative to announce a price change, not necessarily the largest in the industry, but one that accurately reflects changing demand and cost conditions. Other firms in the industry tend to follow suit because it aligns with their interests, considering factors like cost fluctuations and sales performance.

Conversely, **dominant firm price leadership** occurs when a company establishes leadership due to its size, customer loyalty, or cost advantages. This leader may effectively act as a monopolist in its market segment, and other firms follow its pricing lead, often driven by a fear of retaliatory actions from the dominant firm if they undercut prevailing prices.

---

4. Game Theory

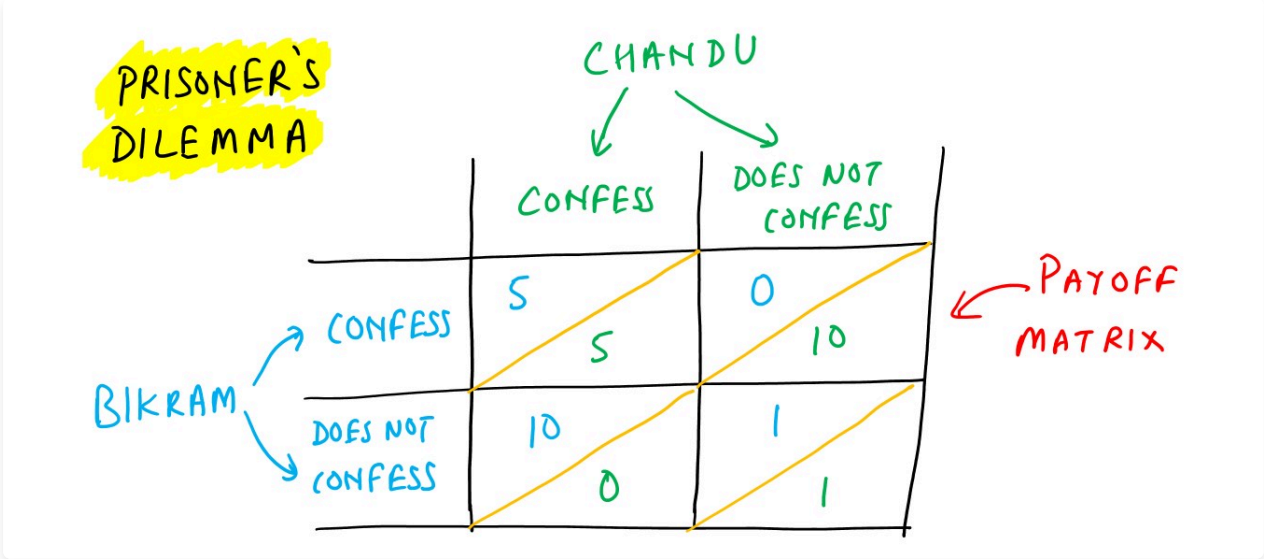
When firms in an oligopoly recognize their interdependence but cannot or choose not to collude, they typically engage in strategic decision-making akin to players in a game. This concept is explored through game theory, which dissects oligopolistic behavior as a sequence of strategic actions and reactions among competing firms. Game theory scrutinizes the choices made by these decision-makers, often referred to as players, whose decisions have a direct impact on each other. In essence, it provides a framework to understand how these firms navigate their competitive environment, anticipating and responding to the actions of their rivals in a dynamic interplay of strategies and counterstrategies.

---

4. Game Theory

To get some feel for game theory, let's work through the prisoner's dilemma, the most widely examined game. It involves two individuals, Bikram and Chandu, who are apprehended near a crime scene and separately interrogated by the police. In this scenario, both individuals are aware that the police have insufficient evidence to prove their guilt without a confession. Each person faces a crucial decision: to either confess, implicating the other, or to remain silent, denying involvement in the crime.

If one of them chooses to confess (squeal), cooperating with the police, he is granted immunity from prosecution and released, while the other person receives a harsh 10-year prison sentence. If both of them remain silent (clam up), they each receive a lighter 1-year sentence due to a technicality. However, if both individuals confess, they both end up serving 5 years in prison.



What will Bikram and Chandu do? The answer depends on the assumptions about their behavior—that is, what strategy each pursues. The figure shows the **payoff matrix** for the prisoner's dilemma. A payoff matrix is a table listing the rewards (or, in this case, the penalties) that the players can expect based on the strategy each pursues. Bikram's choices are shown down the left margin and Chandu's across the top. Each prisoner can either confess or clam up. Notice that the sentence each player receives depends on the strategy he chooses and also on the strategy the other player chooses.

Now, put yourself in Bikram's shoes. If Chandu confesses, and you confess too, you both get 5 years in jail, but if you clam up, you get 10 years and Chandu is free. So, if you think Chandu will confess, you should too. What if you believe Chandu will clam up? If you confess, you do no time, but if you clam up too, you each get 1 year in jail. Thus, if you think Chandu will clam up, you're better off confessing. In short, whatever Chandu does, Bikram is better off confessing. The same holds for Chandu. He is better off confessing, regardless of what Bikram does. So, each has an incentive to confess and both get 5 years in jail. This is called the **dominant-strategy** equilibrium of the game because each player's action does not depend on what he thinks the other player will do.

Thus, we can say that if a player has a dominant strategy in a situation, it will always give at least as high a payoff as any other strategy, whatever other player does. A rational player will always adopt a dominant strategy if one is available.

---

4. Game Theory

The situation becomes more complicated when neither player has a dominant strategy. There is no single equilibrium here. Instead we have to use the concept of a *Nash equilibrium*. This represents an outcome where each player is pursuing their best strategy in response to the best-reply strategy of the other player. Each player should use an equilibrium strategy, one that maximizes each player's expected payoff against the strategy chosen by the other.

It is important to distinguish clearly between a Nash equilibrium and Dominant strategy equilibrium. In a dominant-strategy equilibrium, each player chooses an action that is a best response against any action the other might take. In a Nash equilibrium, each player takes an action that is a best response to the action the other takes.

Both kinds of equilibrium share the essential feature of stability. In equilibrium, there is no second guessing; it is impossible for either side to increase its payoff by unilaterally deviating from its chosen strategy.

The concepts differ in one important respect. When a player has a dominant strategy, there is no circumstance in which doing anything else ever makes sense. The player always should use this strategy. Of course, in many, if not most, competitive situations, players will not have available a single strategy that is dominant. However, as in the market-share competition, there still will be a Nash equilibrium. Here each side's action is a best response against the other's.

In the context of oligopoly firms, the prisoner's dilemma can help explain the challenges that firms face in achieving cooperation.

For example, suppose two firms, A and B, are competing in a market. If both firms cooperate and keep their prices high, they will both benefit from high profits. However, if one firm lowers its prices while the other keeps its prices high, the firm that lowers its prices will gain a competitive advantage, while the other firm will lose market share. If both firms lower their prices, they will both experience lower profits than if they had both cooperated and kept their prices high. In this scenario, both firms face a dilemma. Each firm is tempted to lower its prices to gain a competitive advantage. However, if both firms lower their prices, they will both experience lower profits than if they had both cooperated and kept their prices high.

Even though both firms would benefit from cooperation, they may be tempted to pursue their individual interests by lowering their prices to gain a competitive advantage. This can result in a situation where both firms end up worse off than if they had cooperated.

---

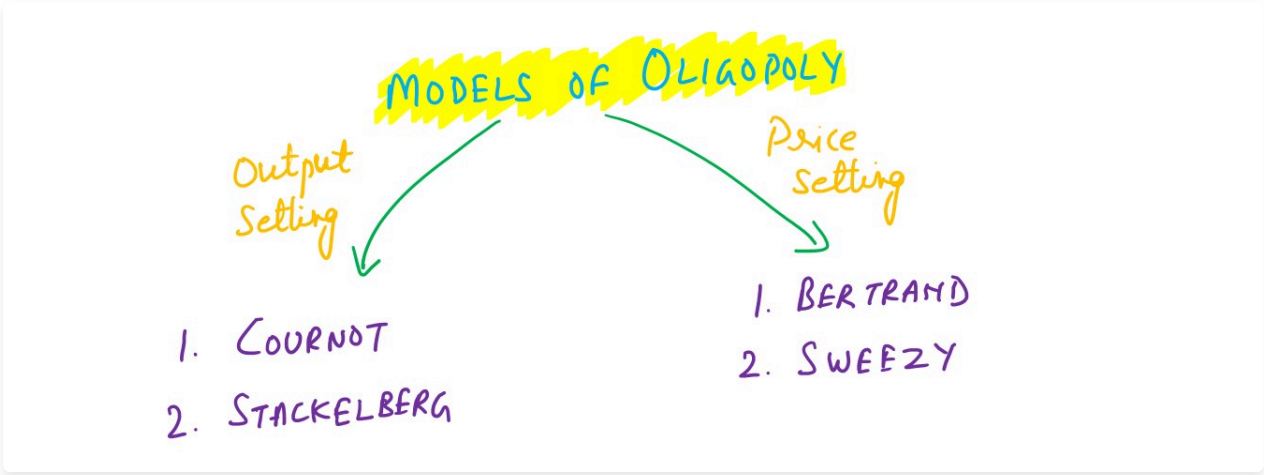
5. Comparison of Oligopoly and Perfect Competition

When comparing oligopoly and perfect competition, it is important to note that while each approach provides insight into the behavior of firms, none provides a complete picture. Here are some key differences to consider:

1. **Price is usually higher under oligopoly**: With fewer competitors in an oligopoly, remaining firms become more interdependent, leading to possible collusion and higher prices. Even without collusion, excess capacity can result in higher prices and lower quantities. In comparison, perfect competition leads to lower prices and higher quantities.
2. **Higher profits under oligopoly:** While perfect competition prevents firms from earning more than a normal profit in the long run, oligopoly can allow for long-run economic profits due to barriers to entry such as economies of scale or brand names. As a result, profits in oligopoly are expected to be higher than under perfect competition.
3. **Behavior depends on barriers to entry:** The lower the barriers to entry in an oligopoly, the more the firms will act like perfect competitors. However, if barriers to entry are high, oligopolists can maintain higher prices and profits.
4. **Greater efficiency in large firms:** Some economists view the higher profit rates in oligopolistic industries as evidence of market power, while others note that the largest firms in these industries tend to earn the highest rate of profit due to greater efficiency arising from economies of scale.

---

1. Introduction



Unlike competitive and monopolistic markets, oligopolistic markets operate depending on either setting the output level or setting the product price. The following are the most common models of oligopoly divided into two groups

1. output-setting group that includes the Cournot and Stackelberg models
2. price-setting group that includes the Bertrand and Sweezy models

---

2. Cournot Model

The Cournot oligopoly model, proposed by the French economist Augustin Cournot, asserts that each firm, in determining its profit-maximizing output level, assumes that the other firm's output will not change and will remain constant. The Cournot model is often illustrated using a two-firm (**duopoly**) market. Each firm believes that its own output strategy does not affect the strategy of its rival(s).

Other firms also act in a similar manner. They attempt to maximize their own profits under the assumption that other firms will continue producing the same amount of output in the next period. In the Cournot model, this pattern continues until reaching the long-run equilibrium point where output and price are stable and neither firm can increase its profits by raising or lowering output. This Cournot equilibrium is also called the **Cournot–Nash equilibrium (CNE)**, since it satisfies the conditions regarding the nature of a Nash equilibrium. The CNE represents the situation where the strategies of the two firms 'match', and there will be no tendency for the firms to change their outputs. At any other pair of outputs there will be a tendency for the firms to change them, since the other firm is not producing what they estimated.

The relationship between an oligopoly firm's profit-maximizing output level and competitor output is called the oligopoly **output–reaction curve** because it shows how oligopoly firms react to competitor production decisions.

---

3. Stackelberg Model

The Stackelberg model of oligopoly is a modification of the Cournot model that takes into account the sequential decision-making process among oligopoly firms (in case of Cournot model it is simultaneous decision making). It was developed by German economist Heinrich von Stackelberg in 1934. It assumes that one firm, the Stackelberg leader, takes the first move by deciding on its output level and then expects the other firm, the Stackelberg follower, to respond to its action. This means that the leader has an advantage over the follower since it can take into account the expected reaction of its competitor in making its own production and pricing decisions. It is often described as first mover advantage model.

In the Stackelberg model, the leader firm's output decision is based on its expectation of the follower's reaction function, which shows how the follower will respond to the leader's output decision. The follower's output decision is then based on its expectation of the leader's output decision, given the leader's assumption about the follower's reaction function. Thus, the parties are interdependent instead of independent.

The Stackelberg model assumes that firms have perfect information about the market, including each other's cost structures, and act rationally to maximize profits. It also assumes that the firms produce homogeneous products and that the leader firm has an advantage over the follower in terms of market power.

Another aspect of oligopoly behavior that the Stackelberg model considers is *price signaling*. This refers to the informal collusion among oligopoly firms through announcing pricing strategies in the hope that competitors will follow suit.

---

4. Bertrand Model

The Bertrand model of oligopoly is a price-setting model that differs from the Cournot and Stackelberg models, which are output-setting models. Developed by French economist Joseph Louis Bertrand in 1883, it assumes that firms in oligopoly markets set prices independently and simultaneously. Each firm sets its own price and assumes that the prices of its rivals remain unchanged. The market quantity is then determined by consumer demand and the given prices.

There are two versions of Bertrand model, when the product is homogenous and when the product is differentiated.

According to Bertrand, when products and production costs are identical, all customers will purchase from the firm selling at the lowest possible price. In case of homogeneous product, a firm can gain market dominance by slightly undercutting its rival's price. However, this triggers a cycle of price cuts until reaching the marginal cost, eliminating any economic profit. This outcome aligns with perfect competition, where no firm earns supernormal profits due to the identical, homogeneous nature of the products offered.

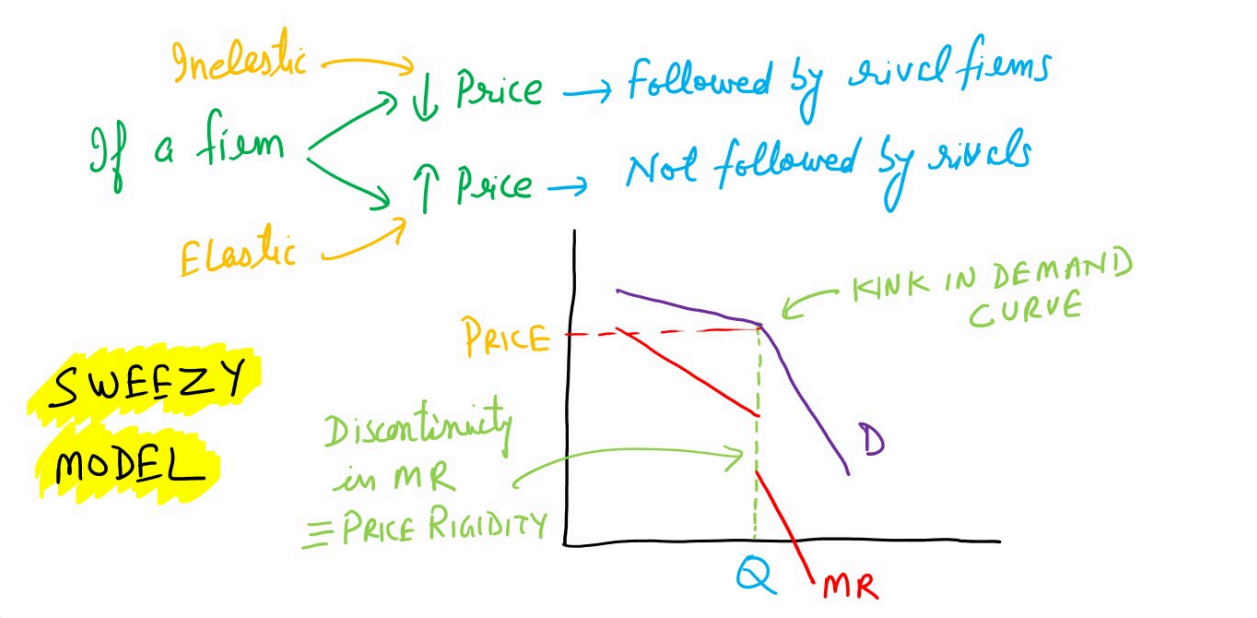
When products are differentiated, as they are in most oligopolistic markets, the analysis of Bertrand competition becomes more complex, and resembles the Cournot model in some respects. First of all, a firm will not lose its whole market if a rival undercuts its price, it will lose only some of its customers. Assuming a two-firm situation again for simplicity, the model is based on each firm having a demand function related both to its own price and to that of its competitor. This corresponds to the Cournot situation where each firm has a demand function related both to its own price and to the output of the competitor.

The relationship between the profit-maximizing price level and competitor price is called the oligopoly **price–reaction curve** because it shows how the oligopoly firm reacts to competitor pricing decisions.

Bertrand equilibrium is reached when no firm can achieve higher profits by charging a different price. The model assumes that firms have perfect information about the market, including their rivals' prices, and act rationally to maximize profits.

---

5. Sweezy Model



This model of oligopoly behavior was developed by, and named after, Harvard University professor of economics Paul Sweezy in 1939. The purpose was to explain price rigidity in the oligopoly market, where it was observed that prices stayed unchanged for a relatively long time. It is also Known as the **Kinked Demand Model**.

Sweezy posited that oligopolistic firms get into a race with each other only when a price decrease is initiated by one firm, but would not engage in a similar race when a price increase is initiated. The interpretation is rather simple. When one firm cuts its price, demand shifts significantly towards the product of that firm, causing the other firms to lose their market share. This is enough to make other firms follow the price cut to maintain their customer base. On the other hand, when one firm increases its price, customers shift away from it and there will be no incentive for other firms to follow. However, the decrease in price would not result in a significant increase in demand for the firm that initiates the price cut, simply because other firms quickly match the cut and mitigate the effect. However, in the case of price increase, almost all customers would shift away making the decrease in demand seem very significant for the firm that initiated the increase. This firm would alone carry the burden of the consequences. Because of the little positive effect of the decrease in price, and the large negative effect of the increase in price, oligopolistic firms tend to prefer to keep the prevailing price that results in price rigidity over a relatively long time.

Thus, the Sweezy theory explains why an established price level tends to remain fixed for extended periods of time in some monopoly markets. Such rigid prices are explained as reflecting a kinked demand curve. A kinked demand curve is a firm demand curve that has different slopes for price increases as compared with price decreases.

Associated with the kink in the demand curve is a point of discontinuity in the marginal revenue curve. As a result, the firm's marginal revenue curve has a gap at the current price–output level, which results in price rigidity.

1. Cost-plus Pricing

Cost-plus pricing is perhaps the most commonly used pricing method by firms. This method involves ascertaining the cost of the product. To this, the amount of profit that a firm desires to earn is added to yield the price of the product. It is also known as **full-cost pricing**.

The concerned cost for this pricing strategy is the average total cost. It comes from the sum of the average variable cost and the average fixed cost. The average fixed cost is the proportionate apportionment of the total fixed cost over the units produced. In case the fixed overheads are marked according to their use in the activities concerned, the method is known as activity based pricing. The profit that has to be added can either be in the form of a percentage mark-up on the cost or an absolute amount.

For example, Suppose that a company sells a product for $1, and that $1 includes all the costs that go into making and marketing the product. The company may then add a percentage on top of that $1 as the "plus" part of cost-plus pricing. That portion of the price is the company's profit.

Depending on the company, the percentage of cost-plus may also include some factor reflecting the current market or economic conditions. If demand is slow, then the cost-plus percentage may be lower in order to lure in customers. On the other hand, if demand for the product is high and economic conditions are good, the cost-plus percentage may be higher as the company feels it can demand a higher price for its product.

---

2. Price Skimming

Price skimming is a mechanism where the seller launches the product in the market at a relatively high price and lowers the price after some period to bring it to a moderate level. It is akin to skimming the cream from milk. As is obvious, this kind of a pricing is more applicable to products that are new or have no competing products with the given specifications. The firm in such cases intends to harness the consumer surplus to its maximum, before the market is flooded with similar such products.

For example, an example of price skimming is DVD players. Initially in 1990s when DVD players were launched the price of a DVD player was $500 and $400. By 2001 the prices were skimmed to less than a $100. By 2004 DVD players were available for as low as $50 or $60.

Another example can be picked up from the computer industry where technology plays a significant role in price skimming. When a new laptop is introduced with enhanced and unique features it is priced quite high. The prices of older laptops now fall as the demand shifts to laptops showcasing new technology.

---

3. Penetration Pricing

Penetration pricing is a pricing method where a firm prices its product fairly low while launching the product and then increases the price to normal levels after a period of time. The technique therefore is just the reverse of the price skimming strategy as discussed earlier, with a difference that here the product may not be new or rather it generally is not new. The firm is new and wishes to occupy some space in the existing product market through lower prices.

By pricing its product low initially the firm plans to attract the existing customers and capture some market share from the existing sellers. This generally happens in market structures where the product differentiation is low and price elasticity of demand is high. In such cases the customers do not mind shifting products, provided they are motivated to do so through reduced prices. The customers can do so easily because with the products nearly close substitutes, they don't associate the product to themselves and the switching costs are relatively low.

For example, television and Internet providers are notorious for their use of penetration pricing — much to the chagrin of consumers who see massive sudden increases in their bills. Comcast/Xfinity, for example, regularly offers low introductory prices such as free or steeply discounted premium channels. At the end of a specified period, the price increases. Most consumers continue paying the higher bill, but some jump to a new provider offering an introductory rate.

A Friday night trip to a video or DVD rental shop was a family tradition across the nation for at least a generation. When Netflix entered the market, it had to convince consumers to wait a day or two to receive their movies. To accomplish this goal, it offered introductory subscription prices as low as a dollar. The pricing strategy was so effective that traditional providers such as Blockbuster soon were edged out of the market.

---

4. Peak Load Pricing

The Peak Load Pricing is the pricing strategy wherein the high price is charged for the goods and services during times when their demand is at peak. In other words, the high price charged during the high demand period is called as the peak load pricing. This type of price discrimination is based on the efficiency, i.e. a firm discriminates on the basis of high usage, high-traffic, high demand times and low demand times. The consumer who purchases the commodity during the high demand period has to pay more as compared to the one who buys during low demand periods.

For example, during summers, the electricity consumption is highest during the daytime as several offices and educational institutes are operational during the day time, called as a **peak-load time**. While the electricity consumption is lowest during the night as all the office establishments and educational institutes are closed by this time, called as **off-peak time**. Thus, a firm will charge a relatively higher price during the daytime as compared to the price charged at night.

---

1. Theories of Firm

During the early 1960s, several economists proposed different theories of firm behavior. These economists, including Simon, Baumol, Marris, Williamson, Berle and Means, Galbraith, and Cyert and March, questioned the validity of the profit maximization hypothesis.

---

1. Theories of Firm

Baumol’s theory of sales maximization argues that business firms aim to maximize sales revenue rather than profits. Baumol believes that most managers seek to maximize sales revenue because they enjoy the discretion to pursue goals other than profit maximization. Baumol's theory suggests that business firms have multiple objectives, making it challenging to identify a single goal that firms commonly pursue.

---

1. Theories of Firm

Williamson’s model of maximization of managerial utility function is a culmination of the managerial utility models. Williamson argues that management is divorced from ownership and that managers have discretionary powers to set the goals of the firm they manage. Managers maximize their own utility function, which includes both quantifiable and unquantifiable variables. Quantifiable variables include managers’ salary, slack earnings, and perks, while unquantifiable variables include power, prestige, job security, status, professional excellence, and discretionary powers to spend money.

---

1. Theories of Firm

Cyert and March viewed large multi-product corporations as a coalition of different but related interest groups, including owners, managers, workers, input suppliers, customers, bankers, and tax authorities. All these groups have their own interests in the corporations and their interests are often in conflict with one another. The top management reconciles these conflicting interests and sets the five main goals of the firm: production goal, inventory goal, sales goal, market share, and profit goal. These goals are determined through a process of continuous bargaining between the coalition groups.

---

1. Theories of Firm

Marris’ theory of firm assumes that the goal that managers of a corporate firm set for themselves is to maximize the firm’s balanced growth rate subject to managerial and financial constraints. In maximizing the firm's growth rate, managers face two constraints: managerial constraints and financial constraints. Managerial constraints arise due to limits to managers’ ability to manage and to achieve optimum efficiency, while financial constraints arise due to a conflict between managers’ own utility function, which they attempt to maximize, and owners’ utility function.

---

1. Theories of Firm

The limit price is the maximum price that existing firms charge with the objective of limiting the number of firms and preventing the entry of new firms to the industry. Limit pricing is a practice of charging a price lower than the profit-maximizing one. Bain's model explains why oligopoly firms maintain their prices over a long period of time at a level that is lower than the price that would maximize their profits.

---

2. Theories of Rent

Rent is a crucial concept in managerial economics, which refers to the payment made by a tenant to a landlord for the use of a factor of production. In economics, rent is not limited to land but also encompasses other factors of production that have a fixed supply in the short run. The theories of rent aim to explain the economic principles behind the payment of rent. We will discuss the Ricardian theory of rent and the concept of quasi-rent.

---

2. Theories of Rent

The Ricardian theory of rent is the earliest known rent theory, propounded by economist David Ricardo. According to this theory, rent is the payment made for the original and indestructible value of the land by the tenant to the landlord. Ricardo believed that rent was a surplus produce that was attributable solely to land as a factor of production. However, modern economists consider rent as an economic surplus that accrues to all factors of production with a fixed supply in the short run.

Ricardo believed that the payment of rent indicated the niggardliness of nature, while French economists known as 'Physiocrats' considered rent as a result of the bounty of nature. The principle of demand and supply is the basis of the Ricardian theory of rent. If the fixed supply of land exceeds the total demand for land in a country, no rent will be paid, just as nothing is paid for the use of air.

---

2. Theories of Rent

The equilibrium price of a factor service has two components: Transfer Earnings and Economic Rent. Transfer earnings or opportunity cost refers to the amount that a factor of production must earn to remain in its present occupation. Economic rent is the excess of the actual earning of a factor over its transfer earning. In other words, economic rent is the factor's actual earning minus its transfer earning.

---

2. Theories of Rent

The concept of quasi-rent was introduced by economist Alfred Marshall. Quasi-rent refers to the short-term earnings of factors of production that have a fixed supply in the short run. In the long run, all inputs are variable, and their supply is elastic. However, in the short run, the supply of certain inputs such as plant and machinery is inelastic. Therefore, the short-term earnings of these factors of production are referred to as quasi-rent.

To illustrate the concept of quasi-rent, let us consider an example. Suppose a firm purchases a machine for Rs 1,00,000 and uses it for five years. After five years, the firm sells the machine for Rs 20,000. The total earnings from the machine during the five-year period were Rs 1,50,000. The transfer earning of the machine was Rs 20,000 (the amount received after selling the machine). Therefore, the economic rent of the machine was Rs 1,30,000 (1,50,000 - 20,000). However, during the five-year period, the machine earned Rs 1,30,000 in quasi-rent, which is the excess of its actual earning over its transfer earning.

---

3. Theories of Profit

Economists define profit as pure profit or economic profit, which is the return over and above the opportunity cost. Pure profit is a residual left over after all contractual costs have been met, including the transfer costs of management, insurable risks, depreciation, and payments to shareholders sufficient to maintain investment at its current level. This residual equals net profit less opportunity costs of management, insurable risk, depreciation of capital, and necessary minimum payments to shareholders that can prevent them from withdrawing their capital from its current use.

---

3. Theories of Profit

According to F.A. Walker, profit is the rent of exceptional abilities that an entrepreneur may possess over the least efficient entrepreneur. Just as rent on land is the difference between the yields of the least fertile and super lands, pure profit is the difference between the receipts of the least efficient entrepreneur and that of those with greater efficiency or managerial ability.

For example, a successful entrepreneur who possesses better managerial skills than the least efficient entrepreneur will earn more profit.

---

3. Theories of Profit

According to J B Clark, profits accrue in a dynamic world, not in a static world. In a static world, there exists absolute freedom of competition, but population and capital are stationary, and there are no inventions. In contrast, in a dynamic world, the factors that remain constant in a static world undergo the process of change. Clark's theory of profit is related to changes in population, capital, production techniques, forms of business organization, and multiplication of consumer's wants.

For example, a dynamic entrepreneur who is capable of taking advantage of changes in the market, promoting their business, expanding sales, and reducing their cost of production will earn more profit.

---

3. Theories of Profit

The risk theory of profit was propounded by F.B. Hawley. He regarded risk-taking as the inevitable accompaniment of dynamic production, and those who take risks have a sound claim to a separate reward, known as **profit**. Thus, profit is simply the price paid by society for assuming business risks.

For example, an entrepreneur who takes on business risks that others are unwilling to take will earn more profit.

---

3. Theories of Profit

Frank H. Knight treated profit as a residual return to uncertainty bearing, not to risk bearing. Knight made a distinction between calculable and incalculable risks. Calculable risks are insurable, while incalculable risks are marked by "uncertainty." Profit arises when an entrepreneur's decision in the area of uncertainty is proved right by subsequent events.

For example, an entrepreneur who makes the right decision in the face of uncertainty will earn more profit.

---

3. Theories of Profit

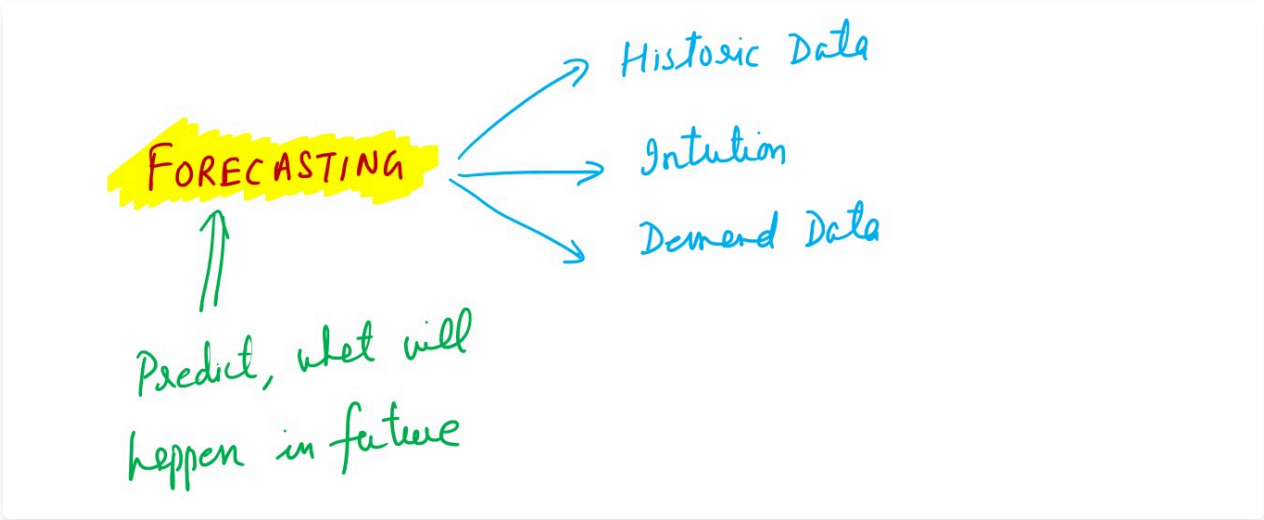
Joseph A. Schumpeter's theory of profit is embedded in his theory of economic development. He assumes a closed, commercially organised, capitalist economy in which private property, division of labour, and free competition prevail, along with a constant population level. Under these conditions of stationary equilibrium, total receipts from the business are exactly equal to the total outlay, and thus there is no profit. Profit can be made by introducing innovations in manufacturing and methods of supplying the goods.

For example, an entrepreneur who introduces a new good, a new method of production, creates or finds a new market, or finds new sources of raw material will earn more profit.

---

1. Introduction

Forecasting is the art and science of predicting future events.



Forecasting may involve taking **historical data** (such as past sales) and projecting them into the future with a mathematical model.

It may be a **subjective or an intuitive** prediction (e.g., "this is a great new product and will sell 20% more than the old one").

It may be based on **demand-driven data**, such as customer plans to purchase, and projecting them into the future.

Or the forecast may involve a combination of these, that is, a mathematical model adjusted by a manager's good judgment.

---

2. Time Frame

The period of forecasting, that is the time range selected for forecasting depends on the purpose for which the forecast is made. The period may vary from one week to some years. Depending upon the period, the forecast can be termed as **Short Range Forecasting**, **Medium Range Forecasting** and **Long Range Forecasting**.

Short Range Forecasting period may be 1 week, 2 weeks or a couple of months. Medium Range Forecasting period may vary from 3 to 6 months. Long Range Forecasting period may vary from 1 year to any period. The objective of above said forecast is naturally different.

Long Range Forecasting
1. to work out expected capital expenditure for future developments 2. to acquire new facilities 3. to determine expected cash flow from sales 4. to plan for future manpower requirements 5. to plan for material requirement 6. to plan for Research and Development
Medium Range Forecasting
1. to determine budgetary control over expenses 2. to determine dividend policy 3. to find and control maintenance expenses 4. to determine schedule of operations 5. to plan for capacity adjustments
Short Range Forecasting
1. to estimate the inventory requirement 2. to provide transport facilities for finished goods 3. to decide work loads for men and machines 4. to find the working capital needed 5. to set-up of production run for the products 6. to fix sales quota 7. to find the required overtime to meet the delivery promises.

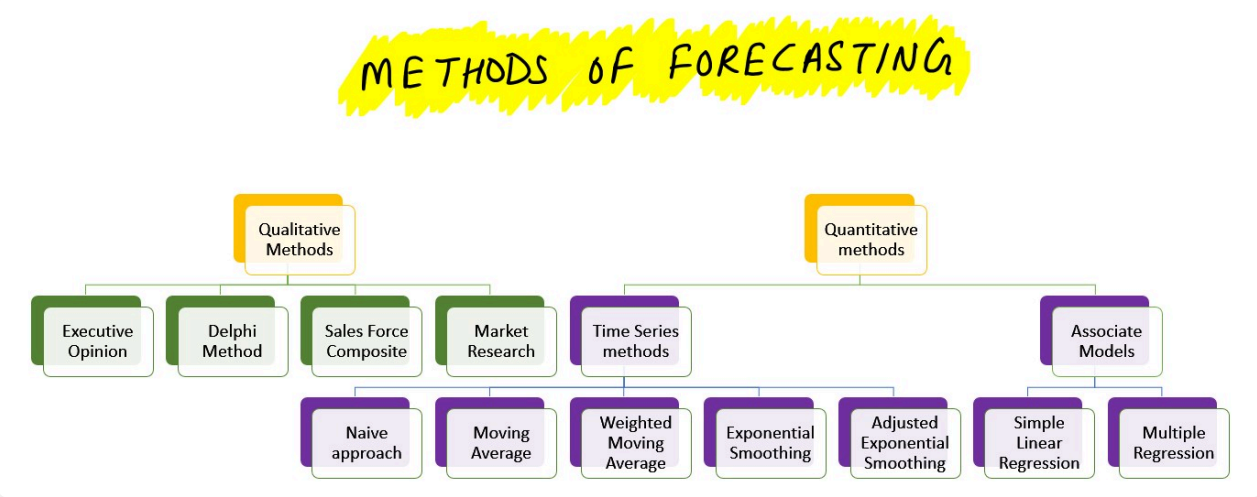
3. Forecasting and Prediction

Key differences between Forecasting and Prediction are given below.

No.	Forecasting	Prediction
1	Forecasting involves the projection of the past into the future.	Prediction reflects management's judgement after taking all available information into account.
2	The forecast involves estimating the level of demand on the basis of factors that generated the demand.	Prediction involves anticipated changes in the future that may or may not have generated the demand.
3	Forecasting is based on a theoretical model.	Prediction may be based on intuition.
4	Forecasting is objective.	Prediction can be biased.
5	The concept used in forecasting is the 'throw ahead' technique which requires a pattern in data.	The concept used in prediction is the 'saying ahead' technique which can be used to predict from random data also.
6	Error analysis is possible.	Error analysis is not possible
7	Forecasting results are replicable.	Prediction is based on unique representations

4. Forecasting Methods

There are two primary categories for forecasting methods: Qualitative and Quantitative.



Qualitative Methods

Qualitative Methods rely on management judgment, expertise, intuition, and opinions to formulate forecasts. Often referred to as "the jury of executive opinion," they are commonly used in long-term strategic planning. Examples include Executive Opinion, Delphi Method, Sales Force Composite, and Market Research.

Quantitative methods

Quantitative methods can be further categorized into Time Series and Associate Models.

Time Series methods, also known as naive models, utilize historical data accumulated over time. These methods assume that past patterns will persist in the future, focusing solely on time as a factor. Techniques under Time Series include Naive Model, Moving Average, Weighted Moving Average, Exponential Smoothing, Adjusted Exponential Smoothing, and Trend Projection.

Associate Models aim to establish a mathematical relationship, typically through regression models, between demand and the influencing factors. They are also called Regression methods or **Causal models**. Examples of Associate Models involve Simple Linear Regression and Multiple Regression.

---

5. Qualitative Forecasts

In some situations, forecasters rely solely on judgment and opinion to make forecasts. If management must have a forecast quickly, there may not be enough time to gather and analyze quantitative data. At other times, especially when political and economic conditions are changing, available data may be obsolete and more up-to-date information might not yet be available. Similarly, the introduction of new products and the redesign of existing products or packaging suffer from the absence of historical data that would be useful in forecasting. In such instances, forecasts are based on executive opinions, consumer surveys, opinions of the sales staff, and opinions of experts.

Executive Opinions

Executive opinion is a forecasting method in which a group of managers meet and collectively develop a forecast. This method is often used for strategic forecasting or forecasting the success of a new product or service. Sometimes it can be used to change an existing forecast to account for unusual events, such as an unusual business cycle or unexpected competition.

Although managers can bring good insights to the forecast, this method has a number of disadvantages. Often the opinion of one person can dominate the forecast if that person has more power than the other members of the group or is very domineering. Think about times when you were part of a group for a course or for your job. Chances are that you experienced situations in which one person's views dominated.

Salesforce Opinions

Members of the sales staff or the customer service staff are often good sources of information because of their direct contact with consumers. In this approach, each salesperson estimates what sales will be in his or her region. These forecasts are then reviewed to ensure that they are realistic. Then they are combined at the district and national levels to reach an overall forecast.

Market Research

Market research is an approach that uses surveys and interviews to determine customer likes, dislikes, and preferences and to identify new-product ideas. Usually, the company hires an outside marketing firm to conduct a market research study. There is a good chance that you were a participant in such a study if someone called you and asked about your product preferences.

Market research can be a good determinant of customer preferences. However, it has a number of shortcomings. One of the most common has to do with how the survey questions are designed. For example, a market research firm may call and ask you to identify which of the following is your favorite hobby: gardening, working on cars, cooking, or playing sports. But maybe none of these is your favorite because you prefer playing the piano or fishing, and these options are not included. This question is poorly designed because it forces you to pick a category that you really don't fit in, which can lead to misinterpretation of the survey results.

Consumer Surveys

This method solicits input from customers or potential customers regarding future purchasing plans. It can help not only in preparing a forecast but also in improving product design and planning for new products. The consumer market survey and sales force composite methods can, however, suffer from overly optimistic forecasts that arise from customer input.

Delphi Method

Another approach is the Delphi method, (Delphi method was originally developed in the early 1950s at the RAND Corporation by Olaf Helmer and Norman Dalkey) an iterative process intended to achieve a consensus forecast. This method involves circulating a series of questionnaires among individuals who possess the knowledge and ability to contribute meaningfully. Responses are kept anonymous, which tends to encourage honest responses and reduces the risk that one person's opinion will prevail. Each new questionnaire is developed using the information extracted from the previous one, thus enlarging the scope of information on which participants can base their judgments.

---

6. Time Series Forecasting

A **time series** refers to a sequential arrangement of observations gathered at consistent intervals, such as hourly, daily, weekly, monthly, or other regular time spans. These observations encompass diverse data types like demand, earnings, accidents, weather patterns, economic indicators like the Consumer Price Index, and various other measurable quantities.

Forecasting techniques based on time-series data operate on the premise that future values within the sequence can be predicted by analyzing past values.

Some examples of time series methods are:

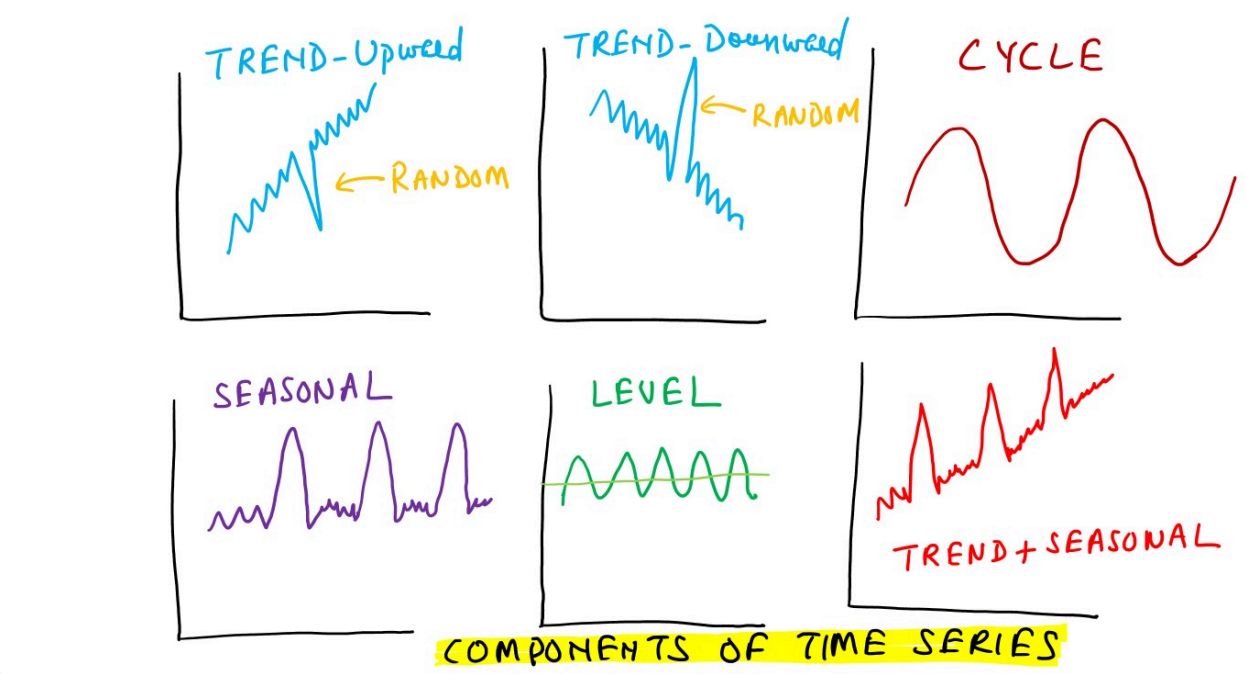
1. *Sales Forecasting*: Predicting future sales figures based on historical sales data to anticipate demand trends and plan inventory.
2. *Financial Markets*: Analyzing historical stock prices to forecast potential price movements and make investment decisions.
3. *Supply Chain Management*: Using past production and delivery data to forecast future demand, aiding in inventory management and logistics planning.

Before delving into the methods of time series analysis, it's crucial to grasp the fundamental components inherent in a time series. These components comprise levels, trends, seasonality, cycles, and variations.

---

7. Components of Time Series

Demand sometimes behaves in a random, irregular way. At other times it exhibits predictable behaviour, with trends or repetitive patterns, which the forecast may reflect. The types of demand behavior are – levels, trends, cycles, and seasonal patterns.

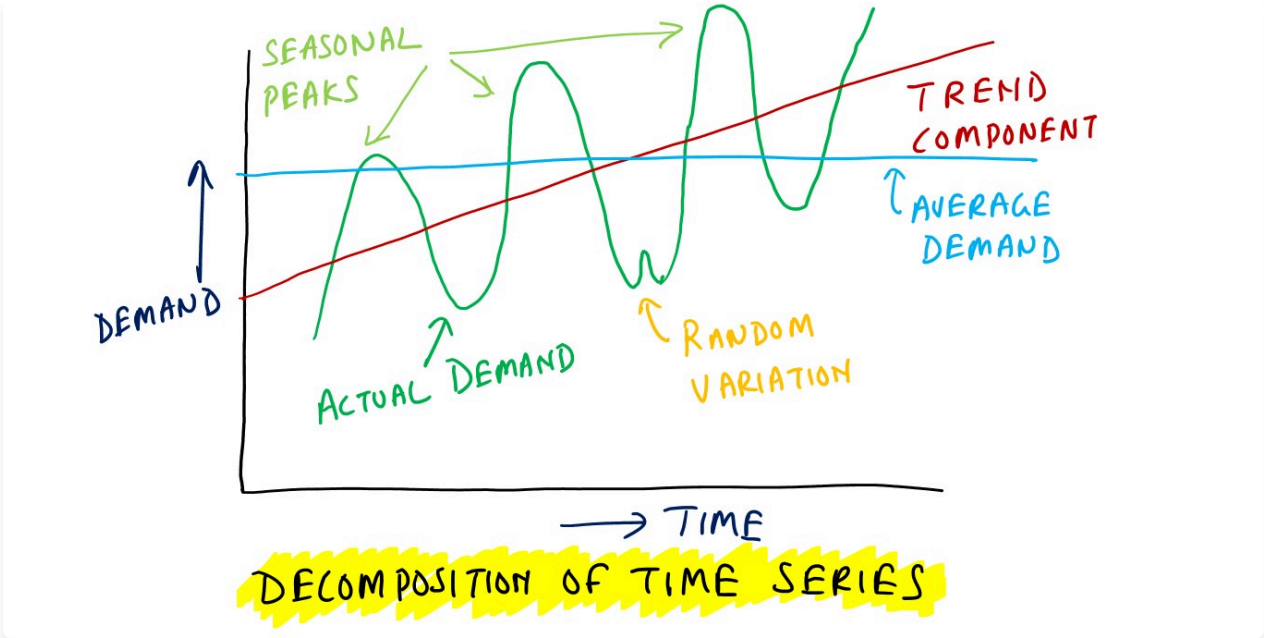


1. Level or horizontal: A level or horizontal pattern exists when data values fluctuate around a constant mean. This is the simplest pattern and the easiest to predict. An example is sales of a product that do not increase or decrease over time. Th is type of pattern is common for products in the mature stage of their life cycle, in which demand is steady and predictable.

2. Trend: When data exhibit an increasing or decreasing pattern over time, we say that they exhibit a trend. The trend can be upward or downward. The simplest type of trend is a straight line, or linear trend.

3. Seasonality: A seasonal pattern is any pattern that regularly repeats itself and is of a constant length. Such seasonality exists when the variable we are trying to forecast is influenced by seasonal factors such as the quarter or month of the year or day of the week. Examples are a retail operation with high sales during October and November because of Diwali or a restaurant with peak sales on Fridays and Saturdays.

4. Cycles: Patterns that are created by economic fluctuations such as those associated with the business cycle are called cycles. These could be recessions, inflation, or even the life cycle of a product. The major distinction between a seasonal pattern and a cyclical pattern is that a cyclical pattern varies in length and magnitude and therefore is much more difficult to forecast than other patterns.



In addition, there will be irregular and random **variations**. *Irregular variations* are due to unusual circumstances such as severe weather conditions, strikes, or a major change in a product or service. They do not reflect typical behaviour, and their inclusion in the series can distort the overall picture. Whenever possible, these should be identified and removed from the data. *Random variations* are residual variations that remain after all other behaviors have been accounted for. The random Variation is also called White Noise.

So, if we look at any forecasted data, we can see that it is composed of the following:
Data = level + trend + seasonality + cycles + random variation

Which can also be written as
Data = pattern + random variation
where pattern = level + trend + seasonality + cycles

---

8. Methods of Time Series

Let us now look at various methods of Time Series forecasting.

---

8. Methods of Time Series

The Naïve Method is one of the simplest forecasting models. It assumes that the next period's forecast is equal to the current period's actual. For example, if your sales were 500 units in January, the naïve method would forecast 500 units for February. It is assumed that there is little change from period to period.

The naïve method can be modified to take trend into account. If we see that our trend is increasing by 10 % and the current period's sales are 100 units, a naïve method with trend would give us current period's sales plus 10 %, which is a forecast of 110 units for the next period. The naïve method can also be used for seasonal data. For example, suppose that we have monthly seasonality and know that sales for last January were 230 units. Using the naïve method, we would forecast sales of 230 units for next January.

One advantage of the naïve method is that it is very simple. It works well when there is little variation from one period to the next. Most of the time we use this method to evaluate the forecast performance of other, more complicated forecasting models. Because the naïve method is simple and effortless, we expect the forecasting model that we are using to perform better than naïve.

---

8. Methods of Time Series

One weakness of the naive method is that the forecast just *traces* the actual data, with a lag of one period; it does not smooth at all. There can be 2 methods of simple averaging.

One of the simplest averaging models is the **simple mean** or **average**. Here the forecast is made by simply taking an average of ALL data points. This model is only good for a level data pattern. As the average becomes based on a larger and larger data set, the random variation and the forecasts become more stable. One of the advantages of this model is that only 2 historical pieces of information need to be carried: the mean itself and the number of observations on which the mean was based.

The **Simple Moving Average** (SMA) is similar to the **simple average** except that we are not taking an average of all the data but are including only 'n' of the most recent periods in the average. As new data become available, the oldest are dropped; the number of observations used to compute the average is kept constant. In this manner, the simple moving average "moves" through time. Like the simple mean, this model is good only for forecasting level data.

The Simple Moving Average forecast can be computed using the following equation:

$F_t = MA_n = \frac{\sum_{i=1}^{n}{A_{t-i}}}{n} = \frac{A_{t-n}+...+A_{t-2}+A_{t-1}}{n}$

where:
F_t = Forecast for time period 't'
MA_n = 'n' period moving average
A_{t-i} = Actual value in period t-i
n = Number of periods (data points in the moving average
For example, MA_3 would refer to a 3-period moving average forecast, and MA_5 would refer to a 5-period moving average forecast.

The disadvantage of the Moving Average method is that it does not react to variations that occur for a reason, such as cycles and seasonal effects. Factors that cause changes are generally ignored. It is basically a "mechanical" method, which reflects historical data in a consistent way. However, the moving average method does have the advantage of being easy to use, quick, and relatively inexpensive. In general, this method can provide a good forecast for the short run, but it should not be pushed too far into the future.

Let us understand this with an example.

Illustration

The Gohana Jalebi wala in Gurgaon, delivers jalebi at home in Gurgaon, Haryana. The owner of the outlet wants to be able to forecast the number of orders that will occur during the next month (i.e., to forecast the demand for deliveries). From records of delivery orders, he has accumulated the following data for the past 10 months, from which it wants to compute simple average, 3-month moving average and 5-month moving average.

SIMPLE AVERAGE

$$= \frac{120+90+100+75+110+50+75+130+110+90}{10} = \frac{950}{10} = 95$$

	Month	Orders	Month	Orders
1	January	120	June 6	50
2	February	90	July 7	75
3	March	100	August 8	130
4	April	75	September 9	110
5	May	110	October 10	90

MOVING AVERAGE

3-MONTHS MOVING AVERAGE

$$MA_3 = \frac{A_{10}+A_9+A_8}{3}$$

$$= \frac{90+110+130}{3} = 110$$

5-MONTHS MOVING AVERAGE

$$MA_5 = \frac{A_{10}+A_9+A_8+A_7+A_6}{5}$$

$$= \frac{90+110+130+75+50}{5}$$

$$= 91$$

8. Methods of Time Series

A weighted average is similar to a moving average, except that it assigns more weight to the most recent values in a time series. For instance, the most recent value might be assigned a weight of 0.40, the next most recent value a weight of .30, the next after that a weight of 0.20, and the next after that a weight of 0.10. Note that the weights must sum to 1.00, and that the heaviest weights are assigned to the most recent values.

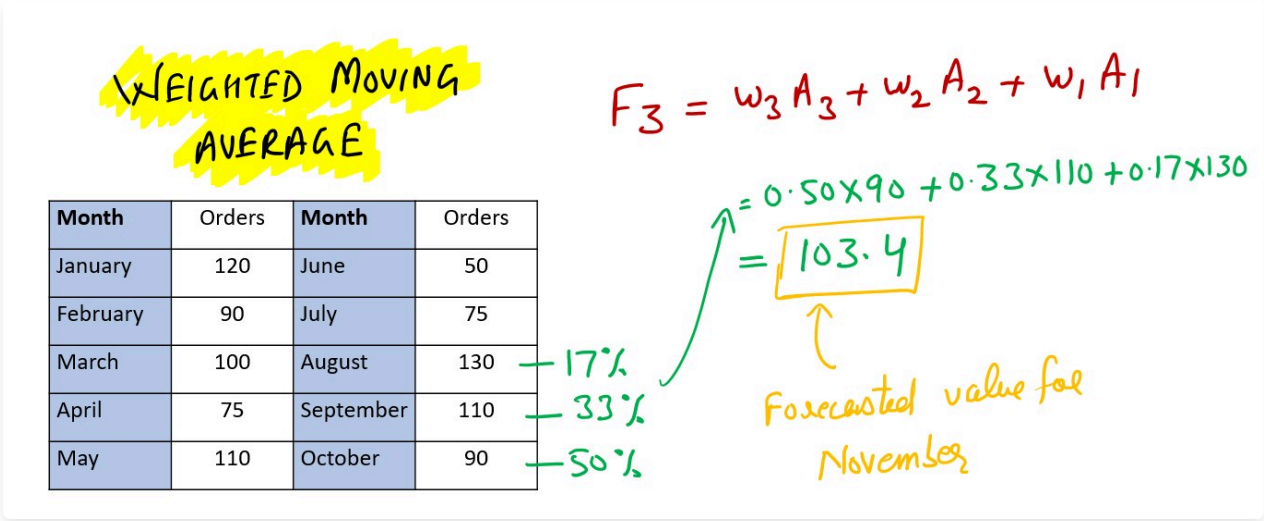
$F_t = w_t(A_t) + w_{t-1}(A_{t-1}) + \ldots + w_{t-n}(A_{t-n})$

where:
w_t= Weight for the period t, w_{t-1}= weight for period t-1, etc.
A_t= Actual value in period, A_{t-1}= Actual value for period t-1, etc.

Let us understand this with an example.

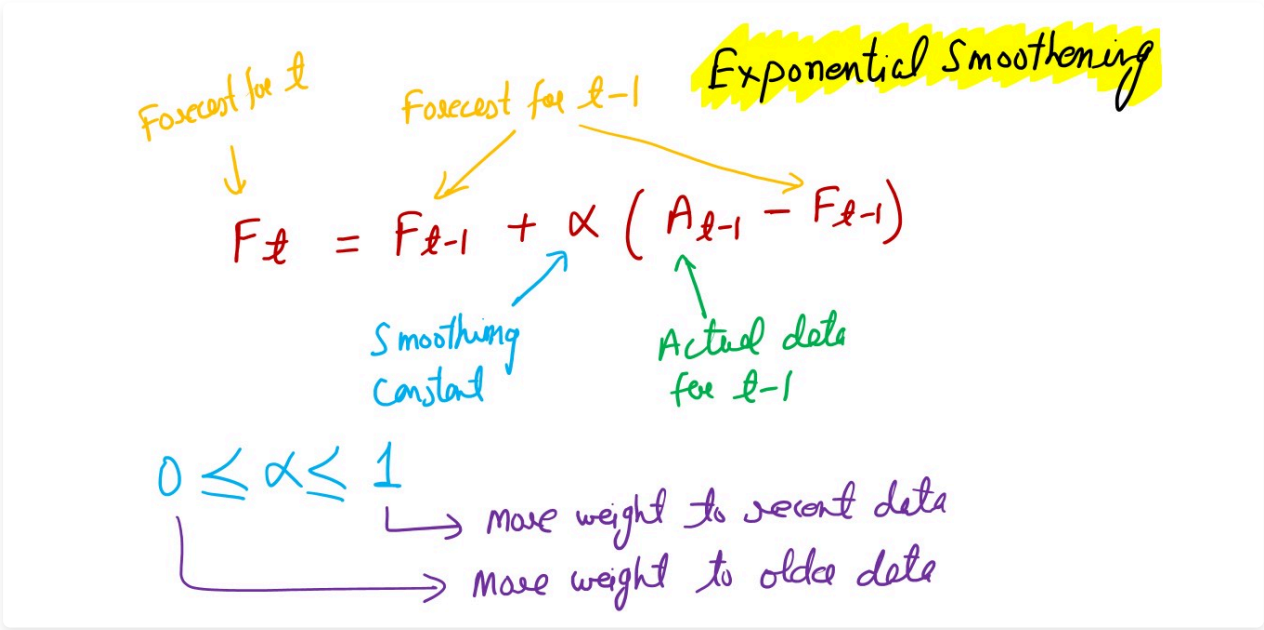
Illustration

The Gohana Jalebi wala in previous question wants to compute a three-month weighted moving average with a weight of 50% for the October data, a weight of 33% for the September data, and a weight of 17% for the August data. These weights reflect the outlet's desire to have the most recent data influence the forecast most strongly.



Month	Orders	Month	Orders
January	120	June	50
February	90	July	75
March	100	August	130
April	75	September	110
May	110	October	90

8. Methods of Time Series



In this method, each new forecast is based on the previous forecast plus a percentage of the difference between that forecast and the actual value of the series at that point. To make a forecast for the next time period, you need 3 pieces of information:

1. The current period's forecast,
2. The current period's actual value
3. The value of a smoothing coefficient, α, which varies between 0 and 1

It weights past data in an exponential manner so that most recent data carry more weight in the moving average.

Next Forecast = Previous Forecast + α(Actual data for previous period– Previous Forecast)

Here (Actual – Previous forecast) represents the forecast error and α is a percentage of the error. More concisely,

$F_t = F_{t-1} + \alpha(A_{t-1} - F_{t-1})$

where:
F_t = Forecast for period t
F_{t-1} = Forecast for the previous period (i.e., period t-1)
α = Smoothing constant (percentage)
A_{t-1} = Actual demand or sales for the previous period

The smoothing constant α represents a percentage of the forecast error. Each new forecast is equal to the previous forecast plus a percentage of the previous error.

For example, suppose the previous forecast was 42 units, actual demand was 40 units, and α was 0.10. The new forecast would be computed as follows:

$F_t = 42 + 0.10(40 - 42) = 41.8$

Then, if the actual demand turns out to be 43, the next forecast would be

$F_t = 41.8 + 0.10(43 - 41.8) = 41.92$

If α is zero, the forecast does not reflect the most recent demand at all. The higher α is, the more sensitive the forecast will be to changes in recent demand, and the smoothing will be less. The closer α is to zero, the greater will be the dampening, or smoothing, effect. As α approaches zero, the forecast will react and adjust more slowly to differences between the actual demand and the forecasted demand.

---

8. Methods of Time Series

A variation of simple exponential smoothing can be used when a time series exhibits a linear trend. It is called **trend-adjusted exponential smoothing** or, sometimes, **double smoothing**, to differentiate it from simple exponential smoothing, which is appropriate only when data vary around an average or have step or gradual changes.

If a series exhibits trend, and simple smoothing is used on it, the forecasts will all lag the trend: If the data are increasing, each forecast will be too low; if decreasing, each forecast will be too high. Unlike a linear trend line, trend-adjusted smoothing has the ability to adjust to changes in trend.

Of course, trend projections are much simpler with a trend line than with trend adjusted forecasts, so a manager must decide which benefits are most important when choosing between these 2 techniques for trend.

Exponential smoothing + Trend adjustment = Adjusted Exponential Smoothing

Forecast with trend = Exponentially smoothened forecast average (F_t) + Exponentially smoothened trend (T_t)

$Ft = \alpha(A_{t-1}) + (1-\alpha)(F_{t-1} + T_{t-1})$

$Tt = \beta(F_t-F_{t-1}) + (1-\beta)T_{t-1}$

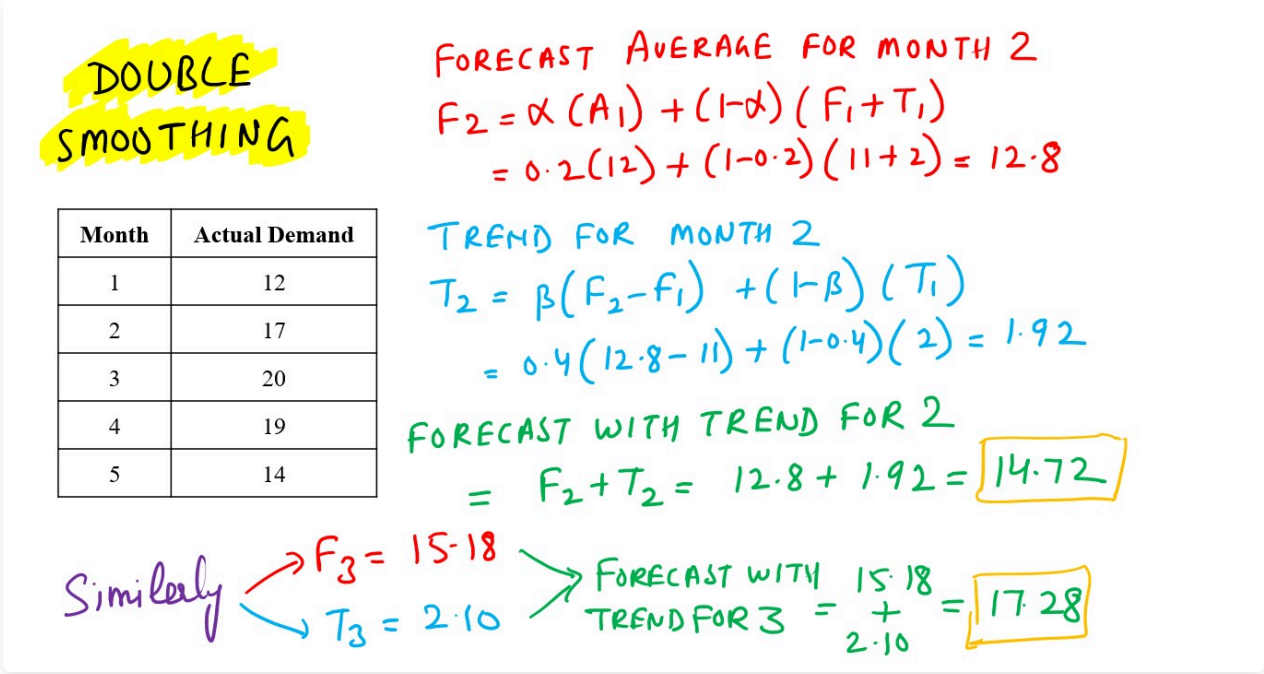
T_{t-1}= Trend estimate in last period

α = Smoothing constant for trend

β = Smoothing constant for average

Illustration

A large Portland manufacturer wants to forecast demand for a piece of pollution-control equipment. A review of past sales, as shown below, indicates that an increasing trend is present. Smoothing constants are assigned the values of α = 0.2 and β=0.4. The firm assumes the initial forecast average for month 1 was 11 units and the trend over that period was 2 units. Calculate forecast including trend for months 2 and 3.



Month	Actual Demand
1	12
2	17
3	20
4	19
5	14

8. Methods of Time Series

An ARIMA model is a class of statistical model for analyzing and forecasting time series data. Along with its development, the authors Box and Jenkins also suggest a process for identifying, estimating, and checking models for a specific time series dataset. This process is now referred to as the **Box-Jenkins Method**.

ARIMA is an acronym that stands for *AutoRegressive Integrated Moving Average*. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

- **AR** – *Autoregression*: A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I** – *Integrated*: The use of differencing of raw observations (i.e., subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
- **MA** – *Moving Average*: A model that uses the dependency between an observation and residual errors from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter.

A standard notation is used of ARIMA (p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

- **p**: The number of lag observations included in the model, also called the lag order.
- **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
- **q**: The size of the moving average window, also called the order of moving average.
