## Content



# Basic Terminologies

## 1. Experiment

- It is basically an activity which I'm trying to do.

> **Let's say I have this mathematical equation**

$a^2 + b^2 + 2ab$

where: $a$ = 3 and $b$ = 4

$3^2 + 4^2 + 2(3)(4)$ = 49

- We are 100% sure that the result of this equation will be 49 only. It cannot be 50 or 48.

This type of experiment is called **Deterministic Experiments** where we can **determine** the exact output, like in this case.

Here are few more examples:

1. Flipping a coin
   - When we flip a coin, there are two possible outcomes: it can land `heads` or `tails`.
   - The outcome can be either of these, each time we perform the experiment.
2. Rolling a six-sided die
   - When we roll the die, the outcome is uncertain, and there are many possible results.
   - The die can land on any of the six faces.
3. Cricket Match
   - Suppose there is a match going on between 2 teams and we don't know what can be the result.
   - Either your team can lose or win. So the outcome is uncertain.

In all of these above examples, we can notice one common thing.

> **Q. Can we determine the outcome of all these experiments?**

No.

Because the outcomes are uncertain. These types of experiments are known as **Probabilistic Experiments**.

Let's continue with the experiment of "Rolling a six sided die" and look at the possible results of an experiment.

### Experiment: Rolling a die

## 2. Outcomes

- Suppose we roll a six sided die and we want to know the possible `Outcomes`.
- We know that we could get any digit out of the 6 digits. So, an outcome could be : {1} or {2} or {3} or {4} or {5} or {6}

## 3. Sample Space

- It is the collection of all the possible outcomes of the experiment.

So the **sample space** for this experiment will be: **{1, 2, 3, 4, 5, 6}**

## 4. Events

We know that sample space for die is {1,2,3,4,5,6}.

If we say,

> **An Even number is rolled / While rolling a die, an even number has occurred**

- Then the possible outcomes will be: `{2, 4, 6}`

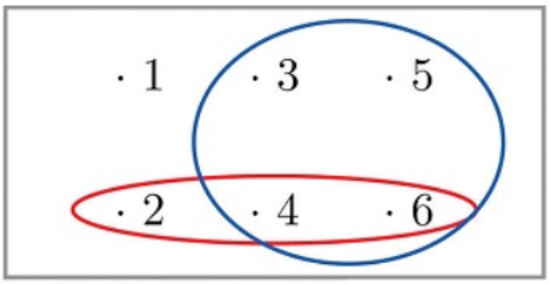
This is known as an **`Event`**.
Any subset of sample space is an event.

- {2, 4, 6} is a subset of sample space.

"**An Even number is rolled**" is an event here and its output is $E$ = {2, 4, 6}, where $E$ denotes an Event.

> **Q1. What are the possible outcomes when a dice is rolled and a number greater than two has occurred?**

- For this Event, outcome will be $E$ = {3, 4, 5, 6}




Here is a graphical representation of a sample space and events

- Here the **sample space** $S$ is represented by a rectangle which is {1, 2, 3, 4, 5, 6}
- **Outcomes** are represented as points within the rectangle which is {1}, {2}, {3}, {4}, {5}, {6}
- **Events** are represented as ovals that enclose the outcomes that compose them.
    - we have two events, $E1$ : {2, 4, 6} which is an event for "Even number is rolled"
    - $E2$ : {3, 4, 5, 6} which is an event for "A number greater than 2 rolled"

Now let's see few experiments.

## Experiment 1: Tossing a single coin

> **Q1. If we toss a single coin then what can be the Possible Outcomes for this experiment?**

- Either we can get **Heads**
- Or we can get **Tails**

Therefore, our outcome becomes: **{H}, {T}**

The **Sample Space** for this experiment will be $S$ **= {H, T}**

> **Based on this sample space, what possible Events can be defined?**

Getting Heads while tossing a coin,

- then our event will be $E$ = {H}

Getting Tails while tossing a coin,

- then our event will be $E$ = {T}

> **Q2. Suppose the given subset is itself {H,T}. Can we define this as an Event or not?**

Yes, It is an event.

- We discussed earlier that any subset of a Sample Space is an Event.
- Also an entire set is a subset of itself so this is a valid event.

> **Q3. So how can we frame this event?**

It is the "**Event of getting Either Heads or Tails**".

> **Q4. Consider the empty set as the given subset denoted by { }. Is it a valid event?**

- We know that, an empty set is a subset of every set. An empty set is therefore a subset of sample space
- It is a valid subset
- So by going with the definition of an Event, we can conclude that this is a valid event.

This can be represented as the "**Event of getting neither Heads nor Tails**".

> **Q5. Is it possible if we toss a coin and get nothing?**

No, it is not possible.

- Therefore, we will have an **Empty set** here

- As we know an empty set is a subset of sample space, therefore it is an Event.

But, the probability of getting a Null Set (No outcome) is Zero.

As it is not possible to toss the coin and don't get any output. we will either gets a head or a tail.

> **Q6. How many subsets can be formed from the sample space?**

There is one formula to find the number of subsets : $2^N$

- where N = number of elements in sample space

For the above experiment, number of elements in the sample sapace is 2 {H,T}, So N = 2

- Therefore the number of subsets will be $2^2 = 4$
- Subsets will be **{ {H}, {T}, {H,T}, { } }**

From this, we can conclude that an empty set is also considered as a valid subset.

## Experiment 2: Tossing 2 coins simultaneously

> **Q1. If we toss 2 coins simultaneously then what can be the Possible Outcomes for this experiment?**

**Explanation :**

- We can get **Heads** and **Heads**
- We can get **Heads** and **Tails**
- We can get **Tails** and **Heads**
- And lastly, we can get **Tails** and **Tails**

Therefore the unique possible **outcomes** will be:

- {HH}, {HT}, {TH}, {TT}

Now, the collection of all the outcomes will be the **Sample Space** :

- {HH, HT, TH, TT}

> **Again, based on this sample space, what are the possible Events that can be defined?**

Let's define few events:

**The Event of getting either one head:**
means one of the outcomes from the 2 tosses should be Head

- {HH, HT, TH}
    - Getting both heads is also valid, getting heads on first coin is also valid and getting heads on the second coin is also valid.

> **Q2. Can we define the above event in a different ways such that the outcome of the event remains the same?**

We can define the above event in 2 other ways:

- **Event of getting at most one tail**
  which means maximum of one Tail only
    - {HH, HT, TH} ,this is the valid outcome

- **Event of getting at least one head**

  which means minimum of one head and maximun 2 as we have 2 coins
    - in this case also {HH, HT, TH} this is the valid outcome

Therefore we can define the above event in 3 different ways

**Another event:**

**Event of getting the same outcomes in both the coins**

- {HH, TT} is the valid outcome as either we get both heads or both tails

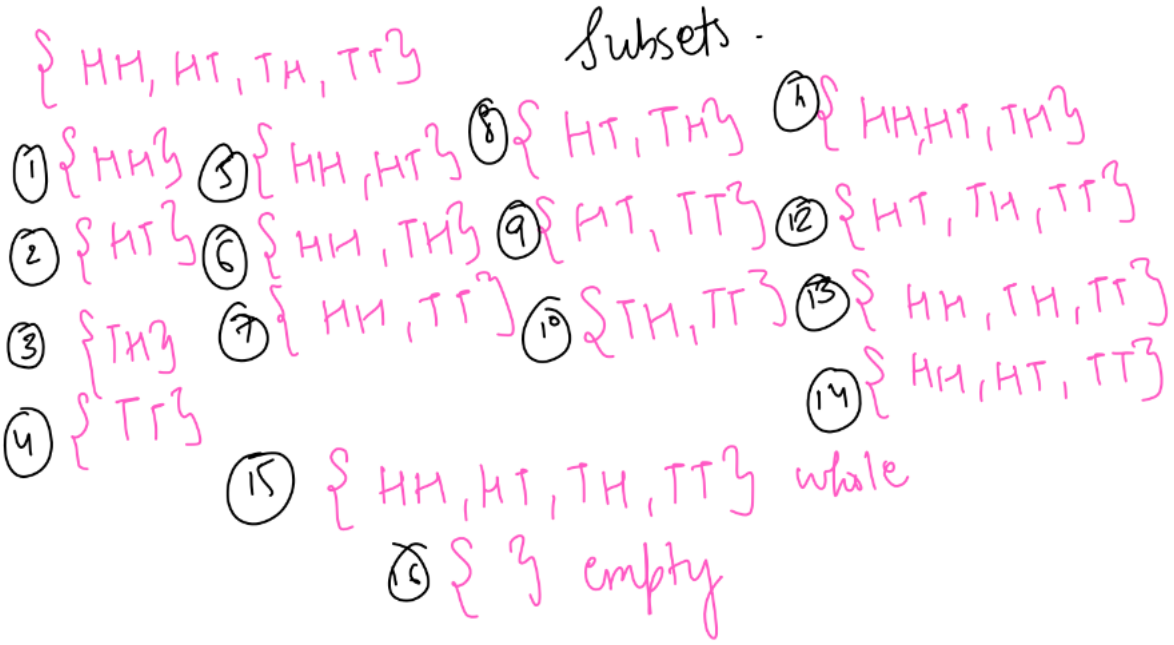
These are some few events, we can define many more.

> **Q3. How many subsets can be formed for this experiment?**

As we know

subset = $2^N$, here $N = 4$

- then number of subsets will be $2^4 = 16$




## Set Operations

Let's recall the Experiment which we defined above i.e. "Rolling a die"

For this experiment we are aware that,

**Sample space** is $\{1, 2, 3, 4, 5, 6\}$

- We can also represent this as a **Universe** or **Universal Set**
- Universal set is the collection of all possible sets

**Now let's define some events**:

- Mohit bets that he will get an odd number
  - So the output of this **Event** will be $A = \{1, 3, 5\}$
- Rakesh bets that he will get either 1, 5 OR 6
  - So the output of this **Event** will be $B = \{1, 5, 6\}$
- Abhishek bets that he will get an Even number
  - So the output of this **Event** will be $C = \{2, 4, 6\}$

Let's discuss few questions with respect to the above events.

### Intersection

> **Q1. In which condition, both Mohit and Rakesh will win their bets?**

They will win their bets when we get a number 1 or 5 on a die.

- If we get number 3, then only Mohit will win his bet as 3 only occurs in his Event.
- And if we get number 6, then only Rakesh will win his bet as 6 only occurs in his event.

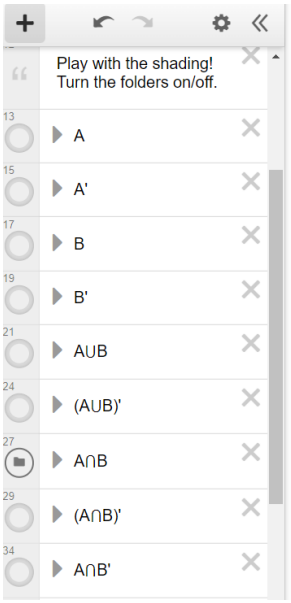
We want a number which occurs in both of their events

- Therefore $\{1, 5\}$ is the possible outcome such that both Mohit and Rakesh will win their bets

This is known as an **Intersection** of two events.

- It is denoted as $A \cap B$
- Intersection means **members belonging to both A `AND` B**
    - If we perform intersection on Events A and B then we will get only those elements that are present in both the events

      i.e. $A \cap B = \{1, 5\}$



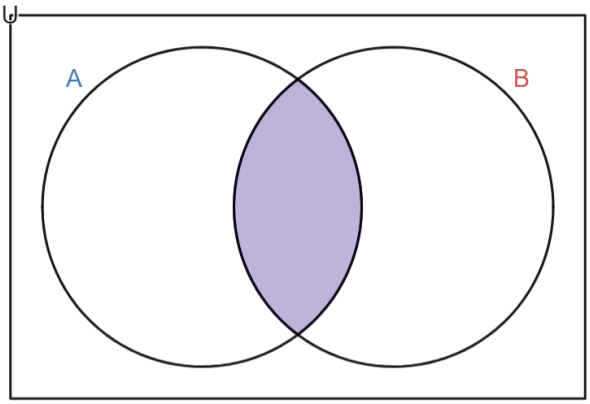




***Image source:** [https://www.desmos.com/calculator/nynlqmtuu2](https://www.desmos.com/calculator/nynlqmtuu2)*

## Union

Now the next question,

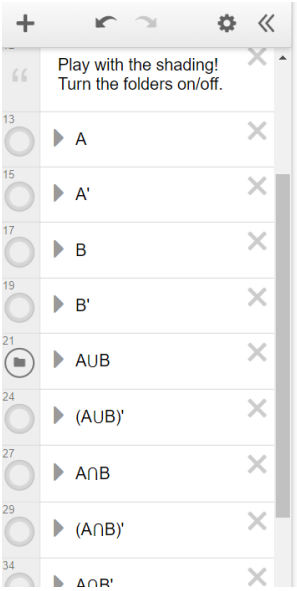
> **Q1. When either Mohit or Rakesh will win their bets?**

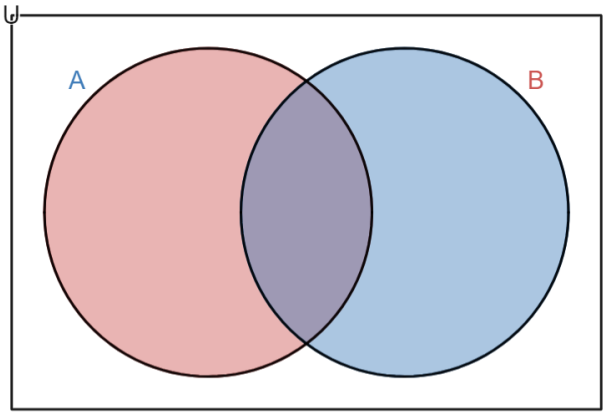
If we get any number out of 1, 3, 5 or 6

- Possible outcomes of this event: $\{1, 3, 5, 6\}$

This is known as **Union** of Two events A and B

- It is denoted by $A \cup B$
- **Union** means **members belonging to either A `OR` B**
- After performing UNION on two events, it'll combine their outcomes together
    - i.e. $A \cup B = \{1, 3, 5, 6\}$




## Complement

> **Q1. When will Mohit lose his bet?**

If we get 2, 4 or 6 because these numbers don't occur in the outcome of Mohit's Event.

- This is known as **complement** of Event A
  - It is denoted by $A'$ or $A^c$

We can define it as the set that contains all the elements Except the elements of A.

- i.e. $A' = \{2, 4, 6\}$

We can also represent it in this way

- $A' = U - A$
  ( represents all the elements minus elements of A)

Similarly we know that Rakesh will lose his bet when we get a 2, 4 or 3

- Hence $B' = \{2, 3, 4\}$

> **Q2. When will Mohit win AND Rakesh lose his bet?**

Mohit will win his bet and Rakesh will lose his bet When we will get a **3** on the dice

- As it is the only outcome which doesn't occur in **Rakesh's event**

There is no new set operation for this, we already know the operations.

To denote this, we will need to use a combination of these operations.

- Essentially we want all elements of A, such that they are not present in B
- This can be written as: $A \cap B' = \{3\}$

> **Q3. Similarly, we can find out when will Rakesh win and Mohit lose his bet**

- $B \cap A' = \{6\}$

> **Q4. What about when BOTH Mohit and Rakesh will lose their bets?**
>
> - $A' \cap B' = \{2, 4\}$

## Mutually Exclusive Events (Disjoint Events)

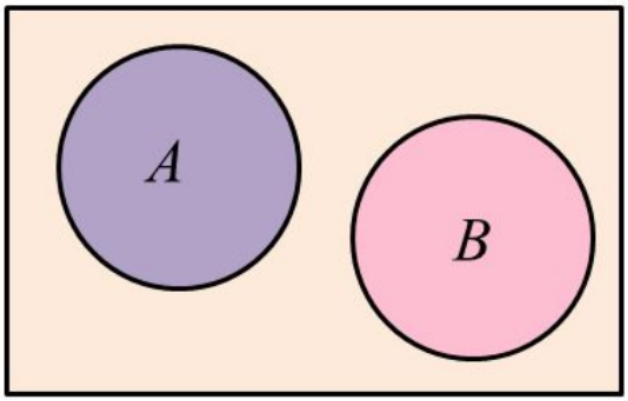
> **Q1. What will be the output of $A \cap C$ ?**

We will have an empty set $\{\ \}$ which can also be represented by $\emptyset$

Because there are no common elements in Set $A$ and Set $C$

Or it implies that **both the events can't occur on the same time** means we can't get an **Even number and a Odd number** at the same time on the dice.

- So, when two events cannot occur at the same time or simultaneously then these types of events are known as `Mutually Exclusive Events` or `Disjoint Events`




$A$ and $B$ are mutually exclusive

## Exhaustive Events

> **Q1. What will be the output of $A \cup B \cup C$?**

Our events are:

- $A = \{1, 3, 5\}$
- $B = \{1, 5, 6\}$
- $C = \{2, 4, 6\}$
  - Therefore $A \cup B \cup C$ = combined elements of Event A, B, C
  - $A \cup B \cup C$ = {1, 2, 3, 4, 5, 6}

We can observe, that all these events when combined together, it covers all the possible outcomes of our experiment i.e **{1, 2, 3, 4, 5, 6}**.

- These types of events are known as `Exhaustive Events`
- Therefore, events `A and B and C are exhaustive events` because, between them, they encompass all the possible outcomes of rolling the die.

## Non Mutually Exclusive Events (Joint Events)

Suppose we define 2 more Events:

- **Event D**: Rolling an even number (2, 4, or 6).
- **Event E**: Rolling a number greater than 3 = (4, 5, or 6).

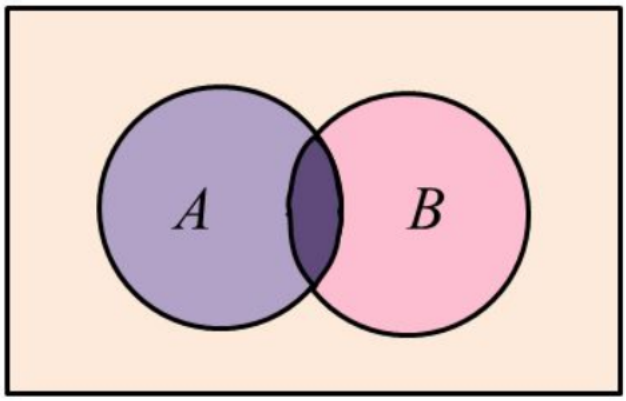
> **Q1. Can we say that Events D and E are mutually exclusive?**

While rolling the die,

we can roll a number that is both even and greater than 3, which means both **events D and E can occur simultaneously**.

- For instance, if the die shows a 4 or a 6, it fulfils the criteria for both events D and E.
- Therefore, events D and E are Not mutually exclusive.

These type of events are known as `non-mutually exclusive` or `joint events`




$A$ and $B$ are not mutually exclusive

## Independent Events

While non-mutually exclusive events allow for overlap, where more than one event can occur, independent events focus on how the occurrence of one event **may or may not affect** the likelihood or outcome of another event

Suppose we have 2 two events:

- **Event A**: Rolling an even number (2, 4, or 6)
- **Event B**: Flipping a coin and getting heads

> **Q1. Are these two events Independent or not?**

YES, these events are `independent Events` because

- The outcome of rolling the die (**Event A**) `does not affect the outcome` of flipping the coin (**Event B**), and vice versa.

They are unrelated events that are occurring independently.

- If we get some number on a die, then it'll not affect the chances of getting Heads or Tails while tossing the coin

These types of events are known as `Independent Events`

And if two events A and B are independent, then the probability of happening of both A and B is:

- $P(A \cap B) = P(A) * P(B)$

In case of Disjoint events, $P(A \cap B)$ **= 0** , as A Intersect B = { }

- **So, if the Events are Independent they cannot be Mutually Exclusive or Disjoint and vice a versa**

In the upcoming lectures, we will see how to derive this formula and also prove this claim.

## How to calculate Probability

Now if I want to calculate the Probability of the particular event let's say event A, then we can calculate using this.

$$Probability = \frac{Outcomes\ in\ set\ A}{Total\ Outcomes\ in\ Entire\ Sample\ Space}$$

Now, let's take a random Experiment whose **outcome** could be **{1} or {2} or {3} or {4} or {5} or {6}**, then the **Sample Space** will be **{1, 2, 3, 4, 5, 6}**

Let's define some events:

1. $A$ **= {2, 4, 6}**

> **Q1. What will be the probability of Event A?**

- By looking into the formula = $\frac{Possible\ oucomes}{Total\ outcomes}$
- Possible outcomes of event A = 3 and total Outcome in sample space = 6

So, $P(A) = \frac{3}{6}$

2. $B$ = **{1, 2}**
   - Similarly Probability of Event B will be $P(B)$ = $\frac{2}{6}$
3. $C$ = **{1, 4, 5, 6}**
   - and Probability of Event C will be $P(C)$ = $\frac{4}{6}$

## Addition Rule

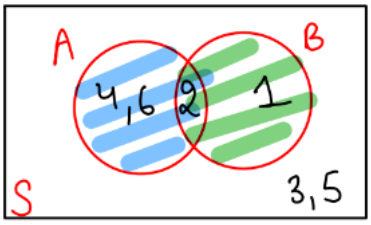
> **Q1. What will be the Probability of $P(A \cup B)$?**

First we need to find $A \cup B$ which is **{1, 2, 4, 6}**

- So by the formula of probability $P(A \cup B)$ will be = $\frac{|A \cup B|}{|S|}$ = $\frac{|\{1,2,4,6\}|}{|\{1,2,3,4,5,6\}|}$ = $\frac{4}{6}$

Where, $|A \cup B|$ = Number of elements(cardinality) of $(A \cup B)$ set,
and $|S|$ = Number of elements in Sample Space

If we want to represent using venn Diagram:




> **Q2. What will be Probability of $P(A \cap B)$?**

$A \cap B$ will be **{2}**

- So by the formula of probability $P(A \cap B)$ will be = $\frac{|\{2\}|}{|\{1,2,3,4,5,6\}|}$ = $\frac{1}{6}$

So by looking into Venn diagram, we observe that AUB means addition of all the elements of **Set A** and **Set B**

- We can also notice in set A we have {2, 4, 6} and in set B we have {1, 2}
- While adding the outcomes of the sets, {2} is occuring twice, which is nothing but $A \cap B$, so we have to subtract it once from our addition, as we want unique outcomes only (Since a set can only have distinct elements).

So the formula for $P(A \cup B)$ can we written as:

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

This is known as **`Addition Rule`**. This is for Joint Events

In case of **Disjoint Events**

- the intersection of $A \cap B = \{\ \}$ so, $P(A \cap B) = 0$
  - therefore, $P(A \cup B) = P(A) + P(B)$

## Experiment 3: Sachin Tendulkar ODI records for India

### Problem Statement:

We have a dataset containing Sachin Tendulkar's ODI cricket career stats, including various performance metrics and the outcomes of matches.

```
!gdown 1TwgJSuiUW8j3_tXsy6B8YwAzAY0tX4Jk
```

```
Downloading...
From: https://drive.google.com/uc?id=1TwgJSuiUW8j3_tXsy6B8YwAzAY0tX4Jk
To: /content/Sachin_ODI.csv
100% 26.4k/26.4k [00:00<00:00, 47.5MB/s]
```

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df_sachin = pd.read_csv("Sachin_ODI.csv")
```

```
df_sachin.head()
```

| | runs | NotOut | mins | bf | fours | sixes | sr | Inns | Opp | Ground | Date | Winner | Won | century |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 13 | 0 | 30 | 15 | 3 | 0 | 86.66 | 1 | New Zealand | Napier | 1995-02-16 | New Zealand | False | False |
| **1** | 37 | 0 | 75 | 51 | 3 | 1 | 72.54 | 2 | South Africa | Hamilton | 1995-02-18 | South Africa | False | False |
| **2** | 47 | 0 | 65 | 40 | 7 | 0 | 117.50 | 2 | Australia | Dunedin | 1995-02-22 | India | True | False |
| **3** | 48 | 0 | 37 | 30 | 9 | 1 | 160.00 | 2 | Bangladesh | Sharjah | 1995-04-05 | India | True | False |
| **4** | 4 | 0 | 13 | 9 | 1 | 0 | 44.44 | 2 | Pakistan | Sharjah | 1995-04-07 | Pakistan | False | False |

Each columns represents different features and each row represents a particular match

```
# shape of the dataset

df_sachin.shape
```

```
(360, 14)
```

**Sample Space** of the experiment consists 360 matches.

## Q1. Out of 360 matches, how many matches did India Won?

```
df_sachin["Won"].value_counts()
```

```
True     184
False    176
Name: Won, dtype: int64
```

Solution 1:

**True = 184**

It indicates that out of the 360 matches, India have won 184 matches and Loose 176 matches

## Q2. A match is randomly chosen, what is the probability that India have won that match?

```
# Probablility of India winning : Possible outcome in event (winning) / Total outcome in sample space = 184/360

184/360
```

```
0.5111111111111111
```

Solution:

If we chose a match at a random from this dataset, then there is a **51% chance** that India have won that match

Let's calculate this using the formula of probability, we know:

probability = $\frac{Possible\ Outcomes\ in\ an\ event}{Total\ Outcomes\ in\ an\ Entire\ Sample\ Space}$

Here we want the possible outcomes of India winning a match (WON = True)

Entire sample space will be our entire dataset

```
# find the rows where India have won and store into new dataframe
df_won=df_sachin.loc[df_sachin["Won"]==True]
```

```
# calculate the number of True values which is our possible outcome
df_won.shape[0]
```

```
184
```

```
# We can also look at the length using len()
len(df_won)
```

- So, probability

  = $\frac{number\ of\ matches\ won}{total\ number\ of\ matches}$

```
prob_winning=len(df_won)/len(df_sachin)
prob_winning
```

```
0.5111111111111111
```

Conclusion: :

If a match is randomly chosen, there is **51%** chance that India have won that match.

## Q3. A match is chosen at a random, what is the probability that Sachin has scored a Century in that match?

Solution 3:

First let's count the **number of centuries**, Sachin has scored

```
# using value_counts()

df_sachin["century"].value_counts()
```

```
False    314
True      46
Name: century, dtype: int64
```

Out of 360 matches, Sachin has scored 46 Centuries.

so, probability of Sachin scoring a century will be:

```
46/360
```

```
0.12777777777777777
```

Similarly we can calculate the probability using second approach too

- We can create a new dataframe which will only contains the data of those matches where Sachin has scored a century

```
df_century=df_sachin.loc[df_sachin["century"]==True]
len(df_century)
```

```
46
```

Then by using the formula of probability we can calculate the probability of Sachin scoring a century

```
prob_century=len(df_century)/len(df_sachin)
prob_century
```

```
0.12777777777777777
```

Conclusion:

If we chose a random match, there is **12.77% chance** that Sachin has scored a century in that match

**Cross Tab:**

Now,

Let's find out how many matches India have won when Sachin has `scored a century` and

How many matches India have won when sachin `didn't score a century`.

> **Q1. Can we achieve this task and obtain all these values at once?**

```
df_sachin[["century","Won"]].value_counts()
```

```
century  Won
False    False    160
         True     154
True     True      30
         False     16
dtype: int64
```

> **Q2. What does these values represent?**

`Century` is representing rows, and `Won` is representing columns

- The first row in the **Century (False)** represents Sachin **didn't score a century**.
- The second row in the **Century (True)** represents Sachin has **scored a century**.

- The first column in the **Won (False)** represents India has **lost the match.**
- The second column in the **Won (True)** represents India has **won the match**

Now, if we take first row and first column (False and False), the value is **160**

- means, India have lost 160 matches when Sachin didn't score a century

If we take first row and second column (False and True), the value is **154**

- means, India have Won 154 matches when Sachin didn't score a century

Similarly, we see that

- India have won only 30 matches when Sachin has scored a century and
- India lost only 16 matches when Sachin has scored a century.

We are able to achieve the task at once using values_counts() but the representation is not great and is not properly aligned.

## Cross Tab and contingency table

> **Q1. Do you remember pivot table from DAV-1 Libraries module?**

- There is a function called `pd.crosstab()` , which accepts parameters **index** and **columns**.

```
pd.crosstab(index=df_sachin["century"],
            columns=df_sachin["Won"],
            margins=True)
```

| Won | False | True | All |
|---|---|---|---|
| century | | | |
| **False** | 160 | 154 | 314 |
| **True** | 16 | 30 | 46 |
| **All** | 176 | 184 | 360 |

What we did using `.valuecounts()` at above, `pd.crosstab()` did the same thing but converted the output into nice tabular format

- **Century** is taken as the **index** and **Won** is taken as **columns**

> **Q2. Now what are the values representing here?**

- These values are giving combinations of **index** and **columns**
  - Wherever **century** is `False` and **Won** is `False` , we get the number of such rows in the dataset or `Frequency` and same for the `True` values

- **160** is the number of rows or frequency representing that number of matches where Sachin didn't score a century (`False`) and India have lost that match (`False`)
- Similarly **154** representing that number of macthes where Sachin didn't score a century (`False`) and India Won that match (`True`)
- **16** represents the number of matches where Sachin has scored a century (`True`) and India have Lost that match (`False`)
- **30** represents the number of matches where Sachin has scored a century (`True`) and India have Won that match (`True`)

This table is also known as `Contingency Table`

> When we do **Margins = True** we get **All**, both in rows and columns, what it represents?

- The values of **All** in a ROW represnts the **Total Value** of each columns (False, True, All)
- The values of **All** in a COLUMN represents the **Total Value** of each rows (False, True, All)

- **176** represents total number of matches we **LOST** (Won -> False)
- **184** represents total number of matches we **WON** (Won -> True)
- **314** represents total number of matches/rows where Sachin **DIDN'T score** a Century (century -> False)
- **346** represents total number of matches/rows where Sachin **scored** a Century (century -> True)
- **360** represents the entire **Sample Space**

We can calculate probabilities using the contingency table.

## Q4. A match is chosen at a random. What is the probability that Sachin has scored a century in that match and India have won that match?

Solution 4:

```
pd.crosstab(index=df_sachin["century"],
            columns=df_sachin["Won"],
            margins=True)
```

| Won | False | True | All |
|---|---|---|---|
| century | | | |
| False | 160 | 154 | 314 |
| True | 16 | 30 | 46 |
| All | 176 | 184 | 360 |

```
# prob of winning and century
# Won -> True, century -> True

30/360
```

```
0.08333333333333333
```

Second Approach:

Can we calculate this probability in a traditional hard coded way?

```
mask = (df_sachin["Won"]==True) & (df_sachin["century"]==True)
mask
```

```
0      False
1      False
2      False
3      False
4      False
       ...
355    False
356    False
357    False
358    False
359    False
Length: 360, dtype: bool
```

```
df_win_and_century=df_sachin.loc[mask ]
len(df_win_and_century)
```

```
30
```

```
prob_win_and_century=len(df_win_and_century)/len(df_sachin)
prob_win_and_century
```

```
0.08333333333333333
```

Conclusion :

There is **8% chance** that Sachin has scored a century and India have won that match if we choose a random match

This tells us, that **contingency table** is more convenient to calculate probabilities rather than hard coded the every single line

## Conclusion of the Problem statement:

Let's have a look how is Sachin's batting can or cannot impact the winning chances of India

1. Out of the **360** matches that Sachin has played, India have **won 184** matches and Loose 176 matches.
2. So, if we choose any match at a random from Sachin's ODI career, there is a **51%** chance that India have won that match.
3. Now, If we choose a random match from Sachin's ODI career, there is **12.77%** chance that Sachin has scored a century in that match.
4. We know if a random match is choosen, there is 12.77% chance that Sachin has scored a century but

   there is **only 8%** chance India have won that match.

   - we can conclude that the chances of India, Winning a match is more when Sachin didn't score a century (what an amazing insight)

Finally,

We can conclude that, if we pick a random match where Sachin played, India's win percentage is 51%. There is 12.77% chance of Sachin scoring a century in that match, and there is only 8% chance that in that match Sachin scores a century as well as India have won that match