Content

- **Problem Solving**
- **Mini Case Study**

Formulas learnt so far

Let's recall all the formulas that we have learned so far,

1. **Conditional probability:** $P[A|B] = \frac{P[A \cap B]}{P[B]}$
2. From conditional probability we will get,

   $P[A \cap B] = P[A|B] \; * \; P[B]$

   which is known as **Multiplication Rule**

1. **Bayes Theorem:** $P[A|B] = \frac{P[B|A] * P[B]}{P[A]}$
2. **Law of total probability:** $P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$
3. **Independent Events:** $P[A \cap B] = P[A] \; * \; P[B]$

Now let's verify one claim.

Claim: If A and B are mutually Exclusive then A and B are not independent.

We know that if A and B are mutually exclusive or Disjoint events:

- $A \cap B = \{\}$
  Note : $A \cap B$ is a null/empty set as A and B can't occur at the same time
- So, $P(A \cap B) = 0$

But in the case of independent events:

- $P(A \cap B) = P(A) * P(B)$ (we just saw above)

In the case of mutually exclusive events $P(A \cap B)$ is not equal to $P(A) * P(B)$, as A and B are not independent.

Therefore, the claim is proven: If A and B are mutually exclusive, then A and B are not independent.

Alternate Method : Using the conditional probability formula:

$$P(A|B) = \frac{P(A{\cap}B)}{P(B)}$$

For Disjoint events:

- $P(A \cap B) = 0$
  - So, $P(A|B) = \frac{0}{P(B)} = 0$

For independent Events:

- $P(A \cap B) = P(A) * P(B)$
  - So, $P(A|B) = \frac{P(A) * P(B)}{P(B)} = P(A)$

As we can see in both the events $P(A|B)$ is different

Hence, we can conclude that :

If A and B are mutually Exclusive then A and B are not independent.

Example: 1

```
In a university, 30% of faculty members are females. Of the
female faculty members, 60% have a PHD. Of the male faculty
members, 40% have a PHD
- What is the probability that a randomly chosen faculty
member is a female and has PHD?
- What is the probability that a randomly chosen faculty
member is a male and has PHD?
- What is the probability that a randomly chosen faculty
member has a PHD?
- What is the probability that a randomly chosen PHD holder
is female?
```

Explanation:

Given,

- Female faculty members = 30%
  - Out of this 30% members, 60% have PHD
- Male faculty members = 100 - 30 = 70%
  - Out of this 70% members, 40% have PHD

Let's define probabilities:

- probability that a randomly chosen faculty member is a female i.e. $P(F) = 0.3$
  - Given that faculty member is a Female, the probability that she has a PHD is i.e $P(phd \mid F) = 0.6$
- probability that a randomly chosen faculty member is a Male i.e. $P(M) = 0.7$
  - Given that faculty member is a Male, the probability that he has a PHD is i.e $P(phd \mid M) = 0.4$

Answering questions:

> **Q1. What is the probability that a randomly chosen faculty member is a female and has PHD?**

We know `AND` means intersection, here we want to find $P(phd \ \cap \ F)$

- Using the formula of conditional probability,

  $P(phd \mid F) = \frac{P(phd \cap F)}{P(F)}$

  So, $P(phd \ \cap \ F) = P(phd \mid F) \ * \ P(F)$

  Adding values into the equation

  - $P(phd \ \cap \ F) = 0.6 * 0.3 = 0.18$

Conclusion:

The probability that a randomly chosen faculty member is a female and has PHD is **0.18**

Similarly,

> **Q2. What is the probability that a randomly chosen faculty member is a male and has PHD?**

- Using the formula of conditional probability,

  $P(phd \mid M) = \frac{P(phd \cap M)}{P(M)}$

  so, $P(phd \ \cap \ M) = P(phd \mid M) \ * \ P(M)$

  Adding values into the equation

  - $P(phd \ \cap \ M) = 0.4 * 0.7 = 0.28$

Conclusion:

The probability that a randomly chosen faculty member is a male and has PHD is **0.28**

> **Q3. What is the probability that a randomly chosen faculty member has a PHD?**

We have 2 approaches to solve this question.

Approach 1:

- Here, we need to find the probability that If I choose a random person, then he/she have a PHD, no matter whether the person is MALE or FEMALE. i.e. $P(phd)$
- We can add $P(phd \ \cap \ F) + P(phd \ \cap \ M)$ as it'll give me $P(phd)$
- $P(phd) = P(phd \ \cap \ F) + P(phd \ \cap \ M)$

  adding values into the equation

  - $P(phd) = 0.18 + 0.28 = 0.46$

Approach 2:

- As we know, we can write $P(phd \ \cap \ F)$ **as a** $P(phd \mid F) \ * \ P(F)$ because, $P(phd \mid F) = \frac{P(phd \cap F)}{P(F)}$

  Here comes the Law of total probability in picture

- For Male also, we can write $P(phd \ \cap \ M)$ **as a** $P(phd \mid M) \ * \ P(M)$

Replacing these values in the equation,

- $P(phd) = [P(phd \mid F) \ * \ P(F)] + [P(phd \mid M) \ * \ P(M)]$
  - $P(phd) = [0.6 * 0.3] + [0.4 * 0.7]$

    = $P(phd) = 0.46$

Conclusion:

The probability that a randomly chosen faculty member has a PHD is **0.46**

> **Q4. What is the probability that a randomly chosen PHD holder is female?**

Here, we are already given that the randomly chosen person is PHD holder and we need to find the probability of this person being Female. We need to find: $P(F \mid phd)$

Using the **formula of conditional probability:**

- $P(F \mid phd) = \frac{P(phd \cap F)}{P(phd)}$

  Replace the $P(phd\ \cap\ F)$ with $P(phd \mid F)\ *\ P(F)$,

  and $P(phd)$ with $[P(phd \mid F)\ *\ P(F)] + [P(phd \mid M)\ *\ P(M)]$

Final forumla will be:

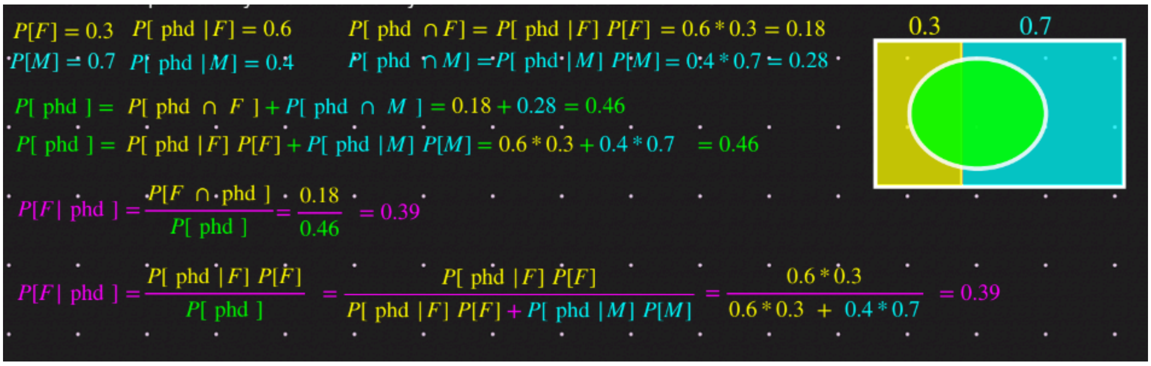
- $P(F \mid phd) = \frac{P(phd \mid F) * P(F)}{[P(phd \mid F) * P(F)] + [P(phd \mid M) * P(M)]}$
  - $P(F \mid phd) = \frac{0.6*0.3}{[0.6*0.3]+[0.4*0.7]}$
  - $P(F \mid phd) = 0.39$

Conclusion:

The probability that a randomly chosen PHD holder is female is **0.39**

There is an alternative approach to solve this question, called **tree based approach**

Let's solve this question with tree based approach.



$P[F] = 0.3 \quad P[\text{ phd } | F] = 0.6 \qquad P[\text{ phd } \cap F] = P[\text{ phd } | F]\ P[F] = 0.6 * 0.3 = 0.18$

$P[M] = 0.7 \quad P[\text{ phd } | M] = 0.4 \qquad P[\text{ phd } \cap M] = P[\text{ phd } | M]\ P[M] = 0.4 * 0.7 = 0.28$

$P[\text{ phd }] = P[\text{ phd } \cap F] + P[\text{ phd } \cap M] = 0.18 + 0.28 = 0.46$

$P[\text{ phd }] = P[\text{ phd } | F]\ P[F] + P[\text{ phd } | M]\ P[M] = 0.6 * 0.3 + 0.4 * 0.7 = 0.46$

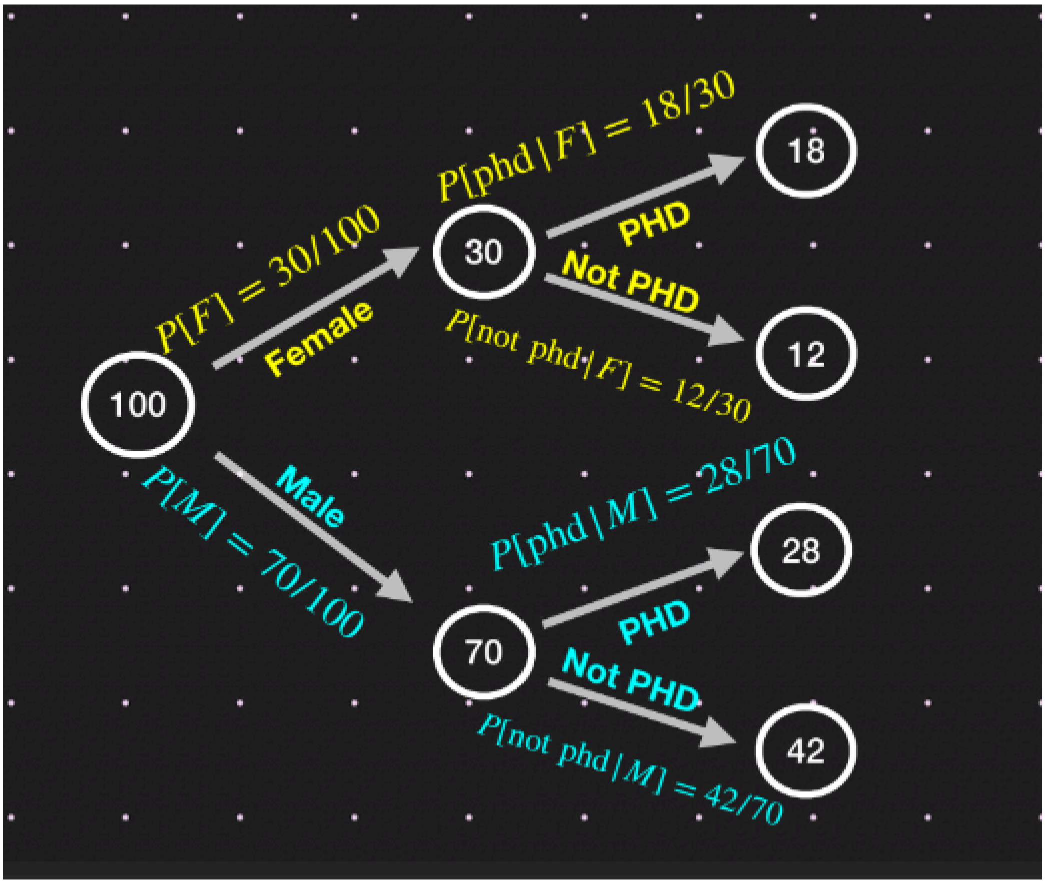
$P[F | \text{ phd }] = \frac{P[F \cap \text{phd }]}{P[\text{ phd }]} = \frac{0.18}{0.46} = 0.39$

$P[F | \text{ phd }] = \frac{P[\text{ phd } | F]\ P[F]}{P[\text{ phd }]} = \frac{P[\text{ phd } | F]\ P[F]}{P[\text{ phd } | F]\ P[F] + P[\text{ phd } | M]\ P[M]} = \frac{0.6 * 0.3}{0.6 * 0.3\ +\ 0.4 * 0.7} = 0.39$

Tree based approach:

Let's assume there are 100 faculty members. Now among these 100 faculty members,

They can be divided into two parts, they can be either male or female.



Explanation of the structure of the Tree:

> **Q1. How many of them are female and how many of them are Male?**

Female : 30% of 100 = 30 (as $P(F) = 0.3$)

We can further seggregate the female part into 2 part:

- Female **AND having** a PHD : 60% of 30 = 18
  - We can represent it as $P(phd \mid F) = 0.6$
- Female **AND NOT having** a PHD : 30 - 18 = 12
  - We can represent it as $P(phd' \mid F) = 1 - P(phd \mid F) = 0.4$

Same for the Male:

Male : 70% of 100 = 70 (as $P(M) = 0.7$)

- Male **AND having** a PHD : 40% of 70 = 28
  - We can represent it as $P(phd \mid M) = 0.4$
- Male **AND NOT having** a PHD : 70 - 28 = 42
  - We can represent it as $P(phd' \mid M) = 1 - P(phd \mid M) = 0.6$

The structure of tree is ready.

Now let's solve the questions

> **Q1. What is the probability that a randomly chosen faculty member is a female and has PHD?**

Let's see how we can easily solve this using tree based approach

We want faculty member and PHD

- From our tree diagram, we can see that there are **18 faculty members who are Female and has PHD**.
    - So $P(F \cap phd)$ = 18/100 = 0.18

      We can observe tha we are getting the same answer but how conviniently we are able to solve this problem with this approach

> **Q2. What is the probability that a randomly chosen faculty member is a male and has PHD?**

Following the same approach as above

- $P(M \cap phd)$ = 28/100 = 0.28

> **Q3. What is the probability that a randomly chosen faculty member has a PHD?**

Here we want to find **total number of faculties having PHD**, it doesn't matter whether the member is male or female

- It will be (18 + 28)/100 = 0.46

> **Q4. What is the probability that a randomly chosen PHD holder is female**?

We have 2 ways to reach the PHD, one through FEMALE and one through MALE

- Now, we need the member **who already has PHD but is a female**.

  It'll be $\frac{18}{18+28}$ = 0.39

> **Q5. What is the probability that a randomly chosen PHD holder is male?
> **

Following the same approach as above

- $P(M \mid phd) = \frac{28}{18+28} = 0.6$

We can see how conviniently and easily we are able to solve all the questions using this Tree based approach

Kerala Flood Case Study

- The dataset contains the monthly rainfall data from years 1901 to 2018 for the Indian state of Kerala.
- It contains the monthly rainfall index of Kerela and also record weather a flood took place that month or not.

In [1]:

```
# Import libraries
import numpy as np
import pandas as pd
```

In [2]:

```
# Read the data
df = pd.read_csv("kerala.csv")
df.head(10)
```

Out[2]:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5
5	KERALA	1906	26.7	7.4	9.9	59.4	160.8	414.9	954.2	442.8	131.2	251.7
6	KERALA	1907	18.8	4.8	55.7	170.8	101.4	770.9	760.4	981.5	225.0	309.7
7	KERALA	1908	8.0	20.8	38.2	102.9	142.6	592.6	902.2	352.9	175.9	253.3
8	KERALA	1909	54.1	11.8	61.3	93.8	473.2	704.7	782.3	258.0	195.4	212.1
9	KERALA	1910	2.7	25.7	23.3	124.5	148.8	680.0	484.1	473.8	248.6	356.6

In [3]:

```
df.shape
```

Out[3]:

```
(118, 16)
```

In [13]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 118 entries, 0 to 117
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   SUBDIVISION      118 non-null    object 
 1   YEAR             118 non-null    int64  
 2   JAN              118 non-null    float64
 3   FEB              118 non-null    float64
 4   MAR              118 non-null    float64
 5   APR              118 non-null    float64
 6   MAY              118 non-null    float64
 7   JUN              118 non-null    float64
 8   JUL              118 non-null    float64
 9   AUG              118 non-null    float64
 10  SEP              118 non-null    float64
 11  OCT              118 non-null    float64
 12  NOV              118 non-null    float64
 13  DEC              118 non-null    float64
 14  ANNUAL_RAINFALL  118 non-null    float64
 15  FLOODS           118 non-null    object 
dtypes: float64(13), int64(1), object(2)
memory usage: 14.9+ KB
```

Let's calculate average rainfall for each month over the years

> **Q. What is the average rainfall for each month over the years**

In [4]:

```
# Calculate the average rainfall for each month
cols = ['JAN', 'FEB', 'MAR', 'APR','MAY','JUN','JUL','AUG','SEP','OCT','NOV

monthly_avg = df[cols].mean()
monthly_avg
```

Out[4]:

```
JAN     12.218644
FEB     15.633898
MAR     36.670339
APR    110.330508
MAY    228.644915
JUN    651.617797
JUL    698.220339
AUG    430.369492
SEP    246.207627
OCT    293.207627
NOV    162.311017
DEC     40.009322
dtype: float64
```

In [5]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

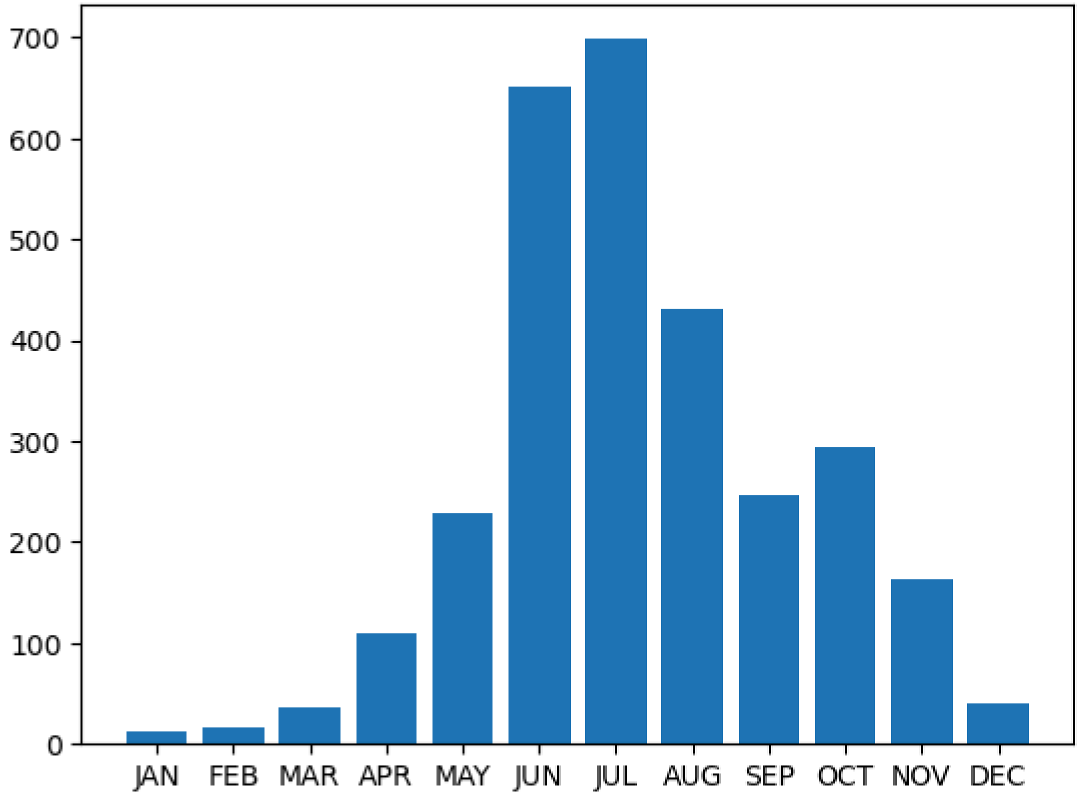
In [6]:

```
x=monthly_avg.index
y=monthly_avg

plt.bar(x,y)
```

Out[6]:

```
<BarContainer object of 12 artists>
```



We can make few **conclusions** here:

- The data reveals significant seasonal variation in rainfall.
    - **June and July** have the **highest average rainfall**, while **January and February** are the driest months
    - The rainfall in **August and September** is still relatively high but begins to decline
    - Surprisingly, **October** has a **higher average rainfall than September**, which may seem counterintuitive.

There are two monsoon seasons in Kerala, **one during Jun-Aug, Other during Oct**.

the important features in this dataset are "JUN", "JUL", "OCT" , "ANNAUL_RAINFALL", "FLOODS"

because in these months only we have seen the peak of the rainfall which can be one of the major source of causing the flood

In [7]:
```
df.columns = [c.replace(' ANNUAL RAINFALL', 'ANNUAL_RAINFALL') for c in df.c

df.head()
```

Out[7]:

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5

In [8]:

```
df.columns
```

Out[8]:

```
Index(['SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JU
L',
       'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'ANNUAL_RAINFALL', 'FLOODS'],
      dtype='object')
```

In [9]:

```
impactful_columns = ['YEAR', 'JUN', 'JUL', 'OCT', 'ANNUAL_RAINFALL', 'FLOODS

impactful_columns
```

Out[9]:

```
['YEAR', 'JUN', 'JUL', 'OCT', 'ANNUAL_RAINFALL', 'FLOODS']
```

Now, I want to label the months column with 0 and 1

- 0: will represents low rainfall
- 1: will represents heavy rainfall

Similarly for "ANNUAL_RAINFALL" column:

- 0: will represents low rainfall in that particular year
- 1: will represents heavy rainfall in that particular year

> **Q. But how much rainfall index is considered as a heavy ranifall?**

One of the parameter is using the **Median** values of these columns.

If their individual **rainfall index value > median value** then it'll we considered as **heavy rainfall** and vice a versa

In [10]:

```
# new dataset containing only impactful columns

data = df[impactful_columns]

data.head()
```

Out[10]:

	YEAR	JUN	JUL	OCT	ANNUAL_RAINFALL	FLOODS
0	1901	824.6	743.0	266.9	3248.6	YES
1	1902	390.9	1205.0	358.4	3326.6	YES
2	1903	558.6	1022.5	354.1	3271.2	YES
3	1904	1098.2	725.5	328.1	3129.7	YES
4	1905	850.2	520.5	383.5	2741.6	NO

In [14]:

```
# Assuming 'data' is a DataFrame with your specified data
threshold_jun = int(data['JUN'].median())
threshold_jul = int(data['JUL'].median())
threshold_oct = int(data['OCT'].median())
threshold_ar = int(data['ANNUAL_RAINFALL'].median())

threshold_jun, threshold_jul, threshold_oct, threshold_ar
```

Out[14]:

```
(625, 691, 284, 2934)
```

In [15]:

```
thresholds = {
    'JUN': 625,
    'JUL': 691,
    'OCT': 284,
    'ANNUAL_RAINFALL': 2934
}

# Convert columns to binary based on thresholds
for col, threshold in thresholds.items():
    data[col] = (data[col] > threshold).astype(int)

data.head()
```

```
/var/folders/zk/yt14z40j2lb2lz548fqr3v9m0000gn/T/ipykernel_19049/443241069.
py:10: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data[col] = (data[col] > threshold).astype(int)
```

Out[15]:

	YEAR	JUN	JUL	OCT	ANNUAL_RAINFALL	FLOODS
0	1901	1	1	0	1	YES
1	1902	0	1	1	1	YES
2	1903	0	1	1	1	YES
3	1904	1	1	1	1	YES
4	1905	1	0	1	0	NO

In [16]:

```
data['FLOODS'].unique()
```

Out[16]:

```
array(['YES', 'NO'], dtype=object)
```

Q1. Calculate the Probability of flood given that rainfall in June is greater than the median june rainfall value (threshold for heavy rainfall)

Question Explanation:

Let A represents : Flood
B represents: heavy rain in June

We need to calculate $P(A|B)$ i.e. $\frac{P(A \cap B)}{P(B)}$

Solution Approach 1:

We can obtain these values using contingency table and put those values into the formula.

Here we need to compare "FLOODS" and "JUN" column.

In [17]:

```python
pd.crosstab(data['JUN'],
            data['FLOODS'],
            margins=True,
            margins_name='Total')
```

Out[17]:

FLOODS	NO	YES	Total
JUN			
0	42	16	58
1	16	44	60
Total	58	60	118

Now, $P(A \cap B)$ = Probability of Flood occuring `AND` heavy rainfall in JUNE

As we know in the contingency table, FLOODS = YES represents that flood has occured and JUN = 1 means heavy rainfall.

We need to check value where FLOODS = YES and JUN = 1 which is **44**

Then by the formula of condititonal probability we can feed this data

In [18]:

```python
# probability of high rainfall in June P(J)
# P(J) = possible outcomes in june having heavy rainfall / total outcomes

P_J = (16+44)/(42+16+16+44)

# now, P(A and B) (Flood = YES and Jun = 1)

P_F_and_J = 44/(42+16+16+44)

#, so our probability of flood occurring given that the high rainfall occure

P_F_J = P_F_and_J / P_J

print(f'P(J) : {P_J}')
print(f'P(F AND J) : {P_F_and_J}')
print(f'P(F|J): {P_F_J}')
```

```
P(J) : 0.5084745762711864
P(F AND J) : 0.3728813559322034
P(F|J): 0.7333333333333334
```

Approach 2: using normalize attribute

Explanation of Normalize attribute:

Rather putting all the values in the formula and then calculate the probability

We can just pass one more attribute in pd.crosstab() function which will divide all values by the sum of values.

- This is the probability only, as in probability we divide `possible outcome / total outcome (sum of all values)`

Parameter is : **normalize = ' '**

- Without this attribute, the contingency table will show the raw counts of occurrences for each combination of variables.
- It will not be normalized, and the values in the table will represent counts.

Here we can pass these strings in this attribute:

normalize='index' or **normalize='columns'** :

- The normalize attribute specifies how the values in the contingency table should be normalized.
    - When set to **'index'**, it calculates conditional probabilities based on rows, treating each row as a separate condition.
    - When set to **'columns'**, it calculates conditional probabilities based on columns, treating each column as the condition we are focusing on.
- This means that each row in the table is divided by the sum of its row, making each row's values sum up to 1, representing conditional probabilities.

Same with the column

In this case:

By setting **normalize='index'**,

- the code calculates conditional probabilities within each row.
- Each value in the table represents the probability of the corresponding event (FLOODS) given the value of 'JUN' in that row.

The row sums up to 1, ensuring that it reflects the conditional probabilities.

In summary,

setting normalize='index' in pd.crosstab allows you to calculate and visualize conditional probabilities based on the specified row variable ('JUN' in this case),

making it easier to assess the impact of one variable on another.

In [19]:
```
pd.crosstab(index = data['JUN'],
            columns = data['FLOODS'],
            margins=True,
            normalize='index')
```

Out[19]:
```
FLOODS        NO       YES
   JUN
     0  0.724138  0.275862
     1  0.266667  0.733333
   All  0.491525  0.508475
```

The values in the table represent the conditional probabilities, where each cell contains the probability of the corresponding outcome (FLOODS) given the condition in June (JUN).

> Then the probability of flood occurring given that the heavy rainfall occured in June will be:

- In the cell at row 1, column 1, the value **0.73333** represents the conditional probability of flooding (FLOODS = YES) given that high rainfall occurred in June (JUN = 1).

Conclusion:

```
  So, there is 73.33% chance of Floods when there is a heavy
rainfall in June
```

As we can see by calculating using formula also, we are getting the same answer as using directly conditional probability using normalize = 'index'

Now, let's jump into the next question

Q2. Given that there is a flooding, calculate the probability that heavy rainfall has occurred in July (more than threshold value)?

Here we want to find $P(July = 1|Flood = YES)$

We are already aware of using formula based approach, so We will solve this using contingency table

Before proceeding,

> **Q. In this question, which string will be passed inside normalize=' ' attribute? 'index' or 'columns'**

In this question, we should normalize the contingency table along the columns

- As we want to find the conditional probability of **high rainfall in July (JUL = 1) given that there was flooding (FLOODS = YES)**,

We want to see how the 'JUL' column behaves when there is flooding.

In [20]:

```
pd.crosstab(index = data['JUL'],
            columns = data['FLOODS'],
            margins=True,
            normalize='columns')
```

Out[20]:

FLOODS	NO	YES	All
JUL			
0	0.655172	0.35	0.5
1	0.344828	0.65	0.5

Conclusion:

The probability that high rainfall occurred in July (JUL = 1) given flooding (FLOODS = YES) is **0.65**.

- **This means that when there is flooding, there is a 65% chance of heavy rainfall in July.**

Q3.Calculate the probability of flood given that june and july rainfall was greater than their median rainfall value?

Solution:

We want to find $P(Flood = Yes|\ june = 1\ and\ Jul = 1)$

Here, we can pass multiple columns in the **pd.crosstab()**

In [21]:

```
pd.crosstab(index = [data['JUN'], data['JUL']],
            columns = data['FLOODS'],
            margins=True,
            normalize='index')
```

Out[21]:

	FLOODS	NO	YES
JUN	JUL		
0	0	0.862069	0.137931
	1	0.586207	0.413793
1	0	0.433333	0.566667
	1	0.100000	0.900000
All		0.491525	0.508475

Conclusion

Frequency (JUN = 1, JUL = 1, FLOODS = YES) = 0.9000000

There is **90%** chance of flood given that heavy rainfall in both june and july

In [ ]: