# ADVANCED ANALYTICS USING STATISTICS– LAB 3

**1. Consider the rolling two 6's in a fair six-sided dice. What is the joint probability of rolling the number five twice in a fair six-sided dice? What is the joint probability of getting a head followed by a tail in a coin toss?**

**Ans=** The joint probability of two independent events A and B occurring is given by the product of their individual probabilities, assuming they are independent events.

For rolling a fair six-sided dice and getting a 5 twice: The probability of rolling a 5 on a fair six-sided dice is 1\6 . Since each roll is independent, the joint probability of rolling a 5 twice is (1\6)^2=1\36.

For the coin toss of getting a head followed by a tail: Assuming a fair coin, the probability of getting a head is 1\2 and the probability of getting a tail is also 1\2 . Since the events are independent, the joint probability of getting a head followed by a tail is (1\2)×(1\2)=1\4.

**2. State the difference between probability density function and probability mass function.**

**Ans=**

**Probability Density Function (PDF):-**

- PDF is associated with continuous random variables.
- It gives the probability that a continuous random variable fall within a particular range of values.
- The area under the PDF curve between two points represents the probability that the variable falls within that interval.

- **Probability Mass Function (PMF):-**

  - PMF is associated with discrete random variables.
  - It gives the probability that a discrete random variable takes on a specific value.
  - For each possible value of the discrete variable, the PMF assigns a probability.
  - The sum of probabilities assigned by the PMF to all possible values of the variable is equal to 1.

**3. Take the Air_Quality dataset and do the following using R code.**

**a)** Do description of data.

**b)** Find if null values are there.

**c)** If present, replace it with mean.

**d)** Find the correlation between each column.

**e)** Calculate correlation coefficient for the two columns pollutant_min and pollutant_max.

**f)** Check whether it is positive or negative correlation.

**g**) Visualize the correlation using scatter plot.

**Ans=** code,

```r
library(readr)
air_quality <- read_csv("Air_Quality.csv")
# a) Description of the data
summary(air_quality)
str(air_quality)

# b) Check for null values
any(is.na(air_quality))

# c) Replace null values with mean for numeric columns
numeric_columns <- sapply(air_quality, is.numeric)
air_quality_numeric <- air_quality[, numeric_columns]
# Replace NA values with mean for each numeric column separately
for (col in colnames(air_quality_numeric)) {
  if (any(is.na(air_quality_numeric[[col]]))) {
    mean_val <- mean(air_quality_numeric[[col]], na.rm = TRUE)
    air_quality_numeric[[col]][is.na(air_quality_numeric[[col]])] <- mean_val
  }
}
# Replace original numeric columns in the dataset with updated numeric columns
air_quality[, numeric_columns] <- air_quality_numeric

# d) Find the correlation between each column
correlation_matrix <- cor(air_quality_numeric)

# e) Calculate correlation coefficient for pollutant_min and pollutant_max
correlation_coefficient <- cor(air_quality$pollutant_min, air_quality$pollutant_max)

# f) Check whether it is positive or negative correlation
if (correlation_coefficient > 0) {
  cat("There is a positive correlation between pollutant_min and pollutant_max.\n")
} else if (correlation_coefficient < 0) {
  cat("There is a negative correlation between pollutant_min and pollutant_max.\n")
} else {
  cat("There is no correlation between pollutant_min and pollutant_max.\n")
}
# g) Visualize the correlation using scatter plot
plot(air_quality$pollutant_min, air_quality$pollutant_max,
     xlab = "pollutant_min", ylab = "pollutant_max",
     main = "Scatter Plot of pollutant_min vs pollutant_max")
```

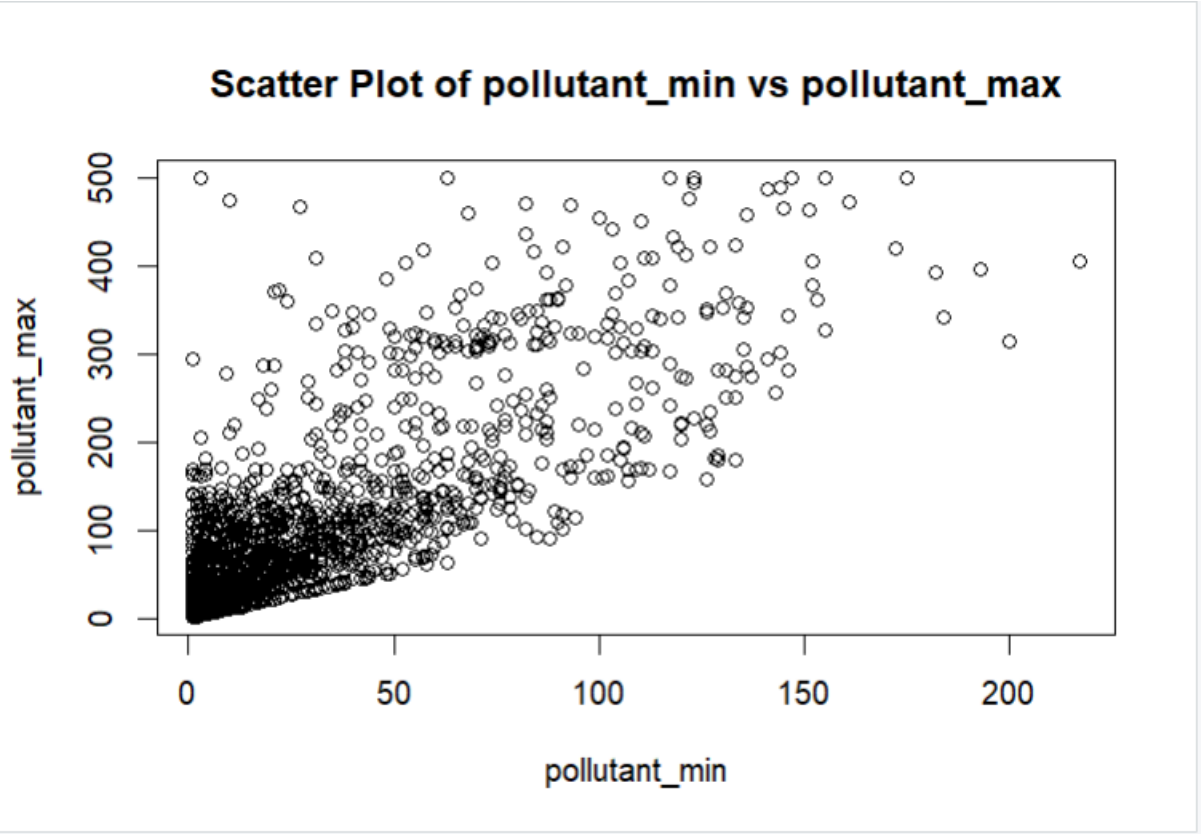
2:43 (Top Level) R Script

```
> library(readr)
> air_quality <- read_csv("Air_Quality.csv")
Rows: 1836 Columns: 10
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (6): country, state, city, station, pollutant_id, last_update
dbl (4): id, pollutant_min, pollutant_max, pollutant_avg

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> summary(air_quality)
       id           country             state               city             station
 Min.   :   1.0   Length:1836        Length:1836        Length:1836        Length:1836
 1st Qu.: 459.8   Class :character   Class :character   Class :character   Class :character
 Median : 918.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 918.5
 3rd Qu.:1377.2
 Max.   :1836.0

 pollutant_id       last_update        pollutant_min     pollutant_max     pollutant_avg
 Length:1836        Length:1836        Min.   :  1.00   Min.   :  1.00   Min.   :  1.0
 Class :character   Class :character   1st Qu.:  5.00   1st Qu.: 21.00   1st Qu.: 12.0
 Mode  :character   Mode  :character   Median : 14.00   Median : 63.00   Median : 31.0
                                       Mean   : 28.41   Mean   : 96.87   Mean   : 54.1
                                       3rd Qu.: 39.00   3rd Qu.:124.00   3rd Qu.: 70.0
                                       Max.   :217.00   Max.   :500.00   Max.   :314.0
                                       NA's   :98       NA's   :98       NA's   :98
> str(air_quality)
spc_tbl_ [1,836 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ id           : num [1:1836] 1 2 3 4 5 6 7 8 9 10 ...
 $ country      : chr [1:1836] "India" "India" "India" "India" ...
 $ state        : chr [1:1836] "Andhra_Pradesh" "Andhra_Pradesh" "Andhra_Pradesh" "Andhra_Pradesh" ...
 $ city         : chr [1:1836] "Amaravati" "Amaravati" "Amaravati" "Amaravati" ...
 $ station      : chr [1:1836] "Secretariat, Amaravati - APPCB" "Secretariat, Amaravati - APPCB" "Secretariat,
Amaravati - APPCB" "Secretariat, Amaravati - APPCB" ...
 $ pollutant_id : chr [1:1836] "PM2.5" "PM10" "NO2" "NH3" ...
 $ last_update  : chr [1:1836] "21-10-2021 01:00" "21-10-2021 01:00" "21-10-2021 01:00" "21-10-2021 01:00" ...
 $ pollutant_min: num [1:1836] 69 82 10 4 16 15 4 47 49 11 ...
 $ pollutant_max: num [1:1836] 109 138 42 5 42 45 82 111 120 44 ...
 $ pollutant_avg: num [1:1836] 86 105 19 4 27 32 42 71 86 23 ...
 - attr(*, "spec")=
  .. cols(
  .. cols(
  ..   id = col_double(),
  ..   country = col_character(),
  ..   state = col_character(),
  ..   city = col_character(),
  ..   station = col_character(),
  ..   pollutant_id = col_character(),
  ..   last_update = col_character(),
  ..   pollutant_min = col_double(),
  ..   pollutant_max = col_double(),
  ..   pollutant_avg = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
> any(is.na(air_quality))
[1] TRUE
> numeric_columns <- sapply(air_quality, is.numeric)
> air_quality_numeric <- air_quality[, numeric_columns]
> for (col in colnames(air_quality_numeric)) {
+   if (any(is.na(air_quality_numeric[[col]]))) {
+     mean_val <- mean(air_quality_numeric[[col]], na.rm = TRUE)
+     air_quality_numeric[[col]][is.na(air_quality_numeric[[col]])] <- mean_val
+   }
+ }
> air_quality[, numeric_columns] <- air_quality_numeric
> correlation_matrix <- cor(air_quality_numeric)
> correlation_coefficient <- cor(air_quality$pollutant_min, air_quality$pollutant_max)
> if (correlation_coefficient > 0) {
+   cat("There is a positive correlation between pollutant_min and pollutant_max.\n")
+ } else if (correlation_coefficient < 0) {
+   cat("There is a negative correlation between pollutant_min and pollutant_max.\n")
+ } else {
+   cat("There is no correlation between pollutant_min and pollutant_max.\n")
+ }
There is a positive correlation between pollutant_min and pollutant_max.
> plot(air_quality$pollutant_min, air_quality$pollutant_max,
+      xlab = "pollutant_min", ylab = "pollutant_max",
+      main = "Scatter Plot of pollutant_min vs pollutant_max")
>
```

Scatter Plot



## 4. Do the following program manually and check with R code.

Suppose we have data on the monthly temperature (in degrees Celsius) and ice cream sales (in thousands of units) for a city over six months:

| Month | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Temperature | 20 | 25 | 18 | 30 | 22 | 28 |
| Ice Cream Sales | 15 | 20 | 12 | 25 | 18 | 22 |

**Hypothesis Formulation:**

- Formulate the null hypothesis ($H_0$) and alternative hypothesis ($H_a$):
  - $H_0$: There is no significant correlation between temperature and ice cream sales ($r = 0$).
  - $H_a$: There is a significant correlation between temperature and ice cream sales ($r \neq 0$).

## Ans=