

PRE-LAB QUESTIONS (PROVIDE BRIEF ANSWERS TO THE FOLLOWING QUESTIONS)

1. Why is multivariate analysis essential in real-world AI problems?

Multivariate analysis is essential because real-world AI problems rarely depend on a single factor. Most systems involve **multiple interacting variables** that jointly influence outcomes. By analyzing these variables together, multivariate methods capture complex relationships, improve predictive accuracy, reduce bias, and enable AI models to make decisions that reflect real-world conditions more reliably.

2. What challenges arise when visualizing high-dimensional data?

Visualizing high-dimensional data is challenging because human perception is limited to two or three dimensions. As dimensionality increases, visualizations can become cluttered, misleading, or difficult to interpret. Key challenges include **loss of information during dimensionality reduction**, overlapping data points, increased cognitive load, and the risk of oversimplifying important relationships.

3. How does correlation analysis support feature selection?

Correlation analysis helps feature selection by identifying **redundant or irrelevant variables**. Features that are highly correlated with each other may provide duplicate information, while weakly correlated features may contribute little to the target variable. Removing such features simplifies the model, reduces overfitting, improves interpretability, and enhances computational efficiency.

4. What are ethical concerns in healthcare data visualization?

Ethical concerns in healthcare data visualization include **patient privacy breaches**, misrepresentation of medical data, and biased interpretations that can affect clinical decisions. Poorly designed visualizations may exaggerate trends, hide uncertainties, or lead to incorrect conclusions, potentially resulting in harmful medical outcomes. Ensuring data anonymization, accuracy, transparency, and fairness is critical.

5. Give examples of multivariate data in AI systems.

Examples of multivariate data in AI systems include:

- **Healthcare AI:** patient age, blood pressure, glucose level, medical history, and genetic data
- **Autonomous vehicles:** speed, GPS coordinates, sensor readings, camera inputs, and weather conditions
- **Recommendation systems:** user preferences, browsing history, ratings, time spent, and demographic data
- **Financial AI:** income, transaction history, credit score, spending patterns, and market indicators

IN-LAB EXERCISE:

OBJECTIVE:

To discover relationships among multiple variables using multivariate visualization.

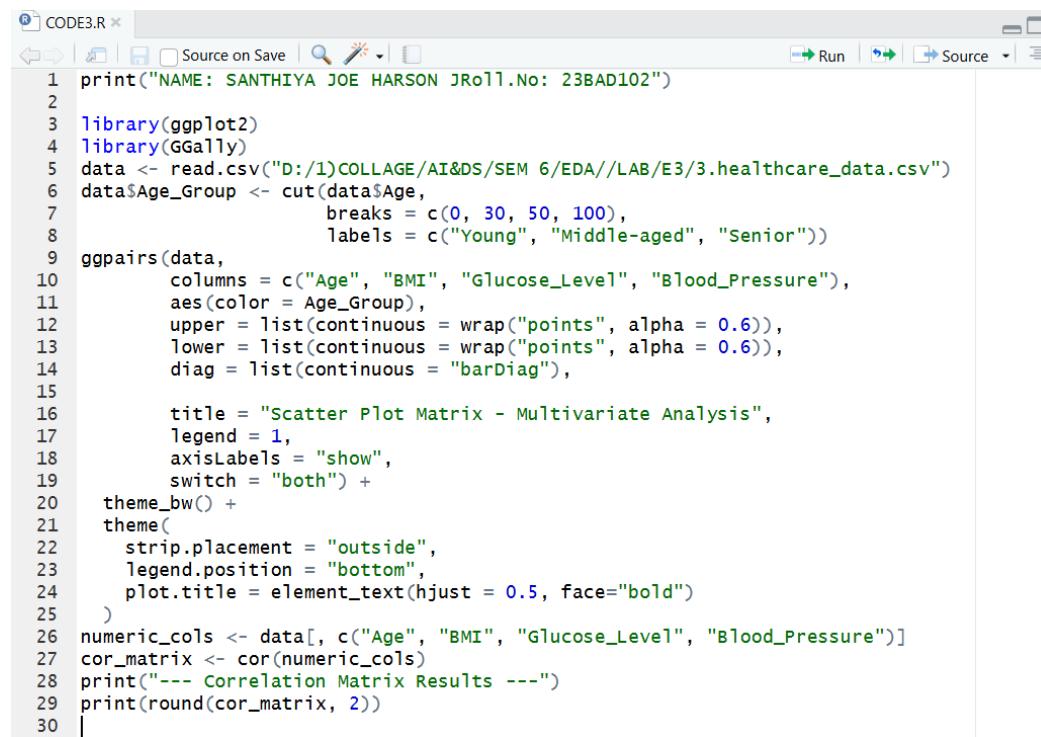
SCENARIO:

A hospital analytics team studies patient health records to identify relationships between age, BMI, glucose levels, and blood pressure for early disease prediction.

IN-LAB TASKS (Using R Language)

- Generate scatter plot matrix
- Apply color encoding for age groups
- Identify correlated health indicators

CODE:



The screenshot shows an RStudio interface with a code editor window titled "CODE3.R". The code is written in R and performs the following steps:

- Prints a name string.
- Loads the ggplot2 and GGally packages.
- Reads a CSV file named "3.healthcare_data.csv" located at "D:/1)COLLAGE/AI&DS/SEM 6/EDA//LAB/E3/3.healthcare_data.csv".
- Cuts the "Age" column into three groups: "Young", "Middle-aged", and "Senior" based on age ranges (0-30, 30-50, 50-100).
- Creates a ggpairs plot for the data. The plot includes:
 - Columns: "Age", "BMI", "Glucose_Level", "Blood_Pressure".
 - Aesthetic: color is mapped to the "Age_Group" factor.
 - Upper panel: continuous points with alpha = 0.6.
 - Lower panel: continuous points with alpha = 0.6.
 - Diagonal panel: barDiag.
 - Title: "Scatter Plot Matrix - Multivariate Analysis".
 - Legend: 1.
 - Axis Labels: "show".
 - Switch: "both".
- Changes the theme to bw() and sets strip placement to outside and legend position to bottom.
- Prints the correlation matrix results.
- Prints the rounded correlation matrix.

OUTPUT:

```
R · R 4.5.2 · D:/1)COLLAGE/AI&DS/SEM 6/EDA/LAB/E3/ ↵
> print("NAME: SANTHIYA JOE HARSON JRoll.No: 23BAD102")
[1] "NAME: SANTHIYA JOE HARSON JRoll.No: 23BAD102"
>
> library(ggplot2)
> library(ggally)
> data <- read.csv("D:/1)COLLAGE/AI&DS/SEM 6/EDA//LAB/E3/3.healthcare_data.csv")
> data$Age_Group <- cut(data$Age,
+                         breaks = c(0, 30, 50, 100),
+                         labels = c("Young", "Middle-aged", "senior"))
> ggpairs(data,
+           columns = c("Age", "BMI", "Glucose_Level", "Blood_Pressure"),
+           aes(color = Age_Group),
+           upper = list(continuous = wrap("points", alpha = 0.6)),
+           lower = list(continuous = wrap("points", alpha = 0.6)),
+           diag = list(continuous = "barDiag"),
+           title = "Scatter Plot Matrix - Multivariate Analysis",
+           legend = 1,
+           axisLabels = "show",
+           switch = "both") +
+   theme_bw() +
+   theme(
+     strip.placement = "outside",
+     legend.position = "bottom",
+     plot.title = element_text(hjust = 0.5, face="bold")
+   )
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
plot: [1, 1] [====>-----] 6% est: 0s
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
plot: [2, 2] [=====>-----] 38% est: 1s
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
plot: [3, 3] [=====>-----] 69% est: 0s
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value `binwidth`.

> numeric_cols <- data[, c("Age", "BMI", "Glucose_Level", "Blood_Pressure")]
> cor_matrix <- cor(numeric_cols)
> print(" --- Correlation Matrix Results --- ")
[1] " --- Correlation Matrix Results --- "
> print(round(cor_matrix, 2))
      Age    BMI Glucose_Level Blood_Pressure
Age  1.00 -0.15    0.11       0.17
BMI -0.15  1.00   -0.12       0.25
Glucose_Level  0.11 -0.12    1.00       0.26
Blood_Pressure  0.17  0.25    0.26       1.00
>
```

The screenshot shows the RStudio interface with four tabs at the top: CODE3.R, data, cor_matrix, and numeric_cols. Below the tabs is a toolbar with icons for back, forward, and search. The main area displays a correlation matrix table with four columns and four rows. The columns are labeled Age, BMI, Glucose_Level, and Blood_Pressure. The rows are also labeled with the same column names. The values in the table are rounded to two decimal places.

	Age	BMI	Glucose_Level	Blood_Pressure
Age	1.00	-0.15	0.11	0.17
BMI	-0.15	1.00	-0.12	0.25
Glucose_Level	0.11	-0.12	1.00	0.26
Blood_Pressure	0.17	0.25	0.26	1.00

CODE3.R data Filter

Patient_ID	Age	Gender	BMI	Blood_Pressure	Glucose_Level	Cholesterol	Disease_Risk	Age_Group
1	3001	31	Male	23.6	104	111	255	Low
2	3002	45	Male	33.9	113	84	241	High
3	3003	65	Female	18.1	127	180	180	High
4	3004	53	Male	21.8	124	102	158	High
5	3005	68	Male	24.2	138	107	200	High
6	3006	33	Male	26.3	158	166	178	Medium
7	3007	45	Female	32.5	151	102	227	Low
8	3008	64	Female	19.5	149	155	189	High
9	3009	46	Male	31.7	139	135	190	Medium
10	3010	28	Male	18.9	98	79	235	Low
11	3011	45	Male	32.3	123	74	160	Low
12	3012	66	Male	18.9	124	192	172	Low
13	3013	41	Female	18.3	90	143	150	Low
14	3014	66	Female	29.8	129	166	195	High
15	3015	49	Female	35.0	153	187	250	High
16	3016	62	Female	33.2	111	107	170	Medium
17	3017	67	Male	27.8	149	116	239	Medium
18	3018	36	Male	33.6	153	178	259	Medium
19	3019	45	Male	18.1	100	84	185	Low
20	3020	55	Female	34.6	103	74	203	Low
21	3021	20	Male	26.3	149	116	236	High
22	3022	27	Female	30.3	119	137	206	Low
23	3023	68	Female	32.0	124	145	150	Low
24	3024	54	Male	30.2	126	199	212	Low
25	3025	71	Male	27.1	94	164	203	Medium

CODE3.R data cor_matrix numeric_cols Filter

	Age	BMI	Glucose_Level	Blood_Pressure
1	31	23.6	111	104
2	45	33.9	84	113
3	65	18.1	180	127
4	53	21.8	102	124
5	68	24.2	107	138
6	33	26.3	166	158
7	45	32.5	102	151
8	64	19.5	155	149
9	46	31.7	135	139
10	28	18.9	79	98
11	45	32.3	74	123
12	66	18.9	192	124
13	41	18.3	143	90
14	66	29.8	166	129
15	49	35.0	187	153
16	62	33.2	107	111



POST-LAB QUESTIONS (PROVIDE BRIEF ANSWERS TO THE FOLLOWING QUESTIONS)

1. Which health parameters show strong correlation?

Based on the correlation matrix, **Glucose Level and Blood Pressure** (≈ 0.26) and **BMI and Blood Pressure** (≈ 0.25) exhibit the strongest positive correlations among the variables analyzed. However, these values remain in the moderate-to-weak range, indicating the absence of strong linear dependence between the health parameters.

2. Why correlation does not imply causation in medical data?

Correlation reflects a statistical association rather than a direct cause-and-effect relationship. In medical datasets, observed correlations may arise due to **confounding factors** such as lifestyle habits, genetic predispositions, medications, or environmental influences. Therefore, an increase in one health parameter does not necessarily cause a change in another.

3. How can these patterns assist predictive healthcare AI?

Recognizing correlation patterns enables healthcare AI systems to **identify meaningful features**, uncover early risk indicators, and model interactions among variables. This supports more accurate disease prediction, particularly for conditions like **diabetes and hypertension**, where multiple health factors contribute simultaneously.

4. What visualization limitations exist for high-dimensional data?

High-dimensional data presents visualization challenges because increasing the number of variables leads to **overlapping plots, visual clutter, and reduced interpretability**. This makes it difficult to clearly identify relationships, trends, or interactions across multiple features at once.

5. How can dimensionality reduction improve visualization?

Dimensionality reduction techniques such as **Principal Component Analysis (PCA)** condense multiple variables into a smaller set of informative components. This approach preserves the most important patterns in the data while enabling clearer **2D or 3D visual representations**, improving interpretability and analytical insight.

ASSESSMENT

Description	Max Marks	Marks Awarded
Pre Lab Exercise	5	
In Lab Exercise	10	
Post Lab Exercise	5	
Viva	10	
Total	30	
Faculty Signature		