Name: ujwal sahu	Batch: B	Roll no : 34
Subject: R Programming	EXPERIMENT NO: 8	SY-BVOC-SEM IV

Title: Implement PCA

Tools: R studio

Theory: PCA:

Principal Component Analysis (PCA) is a powerful statistical technique used for dimensionality reduction, which simplifies complex datasets by transforming them into a new set of uncorrelated variables called principal components. These components are ordered such that the first few retain most of the variation present in the original dataset. PCA is especially useful when working with high-dimensional data, where it becomes difficult to visualize or model patterns effectively. The PCA process begins by centering and scaling the data, ensuring that all variables contribute equally regardless of their original scale. Then, the algorithm computes the covariance matrix to understand relationships between variables. The eigenvectors and eigenvalues of this matrix are then calculated. The eigenvectors represent the directions (principal components), and the eigenvalues measure the amount of variance captured along each direction. By selecting the top ‘k’ principal components, PCA helps retain most of the original variability with fewer dimensions. In R, PCA is implemented using functions like prcomp() or princomp(), which internally apply linear algebra techniques (such as SVD) to extract principal components. Visual tools like scree plots and biplots help interpret the results, making PCA both a statistical and exploratory analysis tool. Applications of PCA include image compression, pattern recognition, data visualization, and noise reduction, making it a core technique in data analysis, machine learning, and scientific computing.

Implementation Steps:

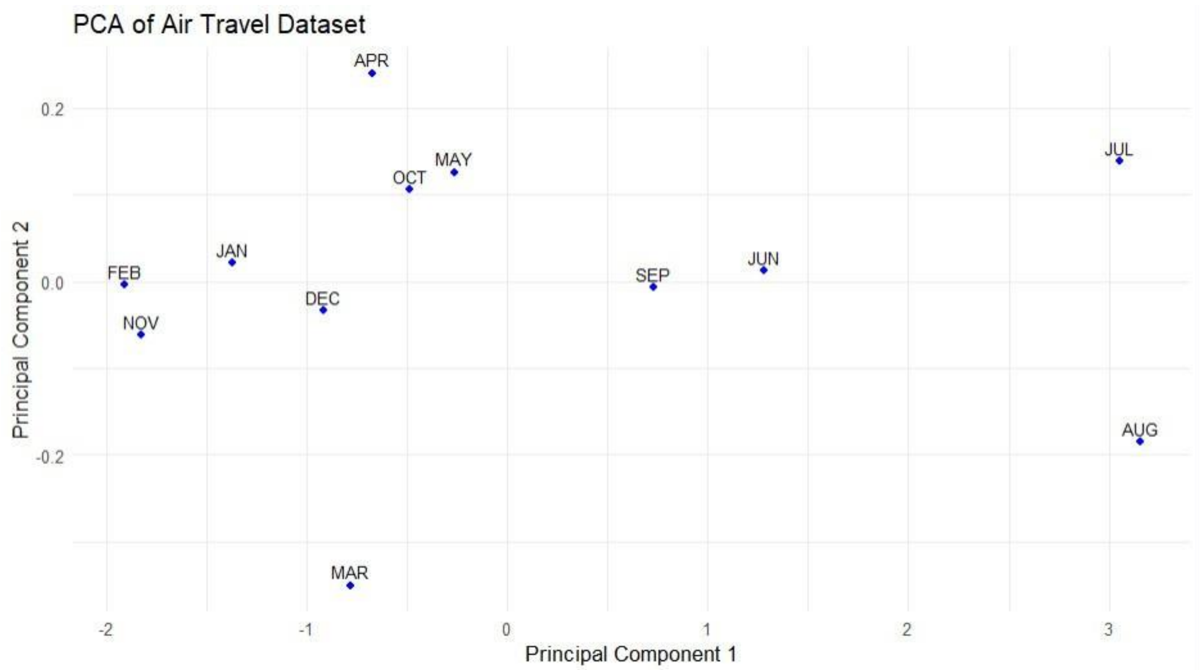
```
# Load required libraries
library(ggplot2)
library(readr)
```

```
> # Step 1: Load the dataset from the URL
> url <- "https://people.sc.fsu.edu/~jburkardt/data/csv/airtravel.csv"
> airtravel <- read_csv(url)
Rows: 12 Columns: 4
── Column specification ────────────────────────────────────────
Delimiter: ","
chr (1): Month
dbl (3): 1958, 1959, 1960
```

```
> # Step 2: Inspect the dataset
> print(head(airtravel))
# A tibble: 6 × 4
  Month `1958` `1959` `1960`
  <chr>  <dbl>  <dbl>  <dbl>
1 JAN      340    360    417
2 FEB      318    342    391
3 MAR      362    406    419
4 APR      348    396    461
5 MAY      363    420    472
6 JUN      435    472    535

> # Step 3: Data Cleaning and Preparation
> # Convert the dataset to numeric format (excluding the first column if it's a categorical variable)
> airtravel_numeric <- as.data.frame(lapply(airtravel[,-1], as.numeric))
> # Handle missing values (if any)
> airtravel_numeric[is.na(airtravel_numeric)] <- 0  # Replace NA with 0
> # Step 4: Perform PCA
> pca_result <- prcomp(airtravel_numeric, center = TRUE, scale. = TRUE)
> # Step 5: Print PCA summary
> summary(pca_result)
Importance of components:
                          PC1     PC2     PC3
Standard deviation     1.7194 0.15558 0.13901
Proportion of Variance 0.9855 0.00807 0.00644
Cumulative Proportion  0.9855 0.99356 1.00000

> # Step 6: Visualize PCA results
> pca_data <- data.frame(pca_result$x)
> pca_data$Month <- airtravel$Month  # Retaining the original month column
> # Scatter plot of the first two principal components
> ggplot(pca_data, aes(x = PC1, y = PC2, label = Month)) +
+   geom_point(color = "blue") +
+   geom_text(vjust = -0.5, size = 3) +
+   labs(title = "PCA of Air Travel Dataset",
+        x = "Principal Component 1",
+        y = "Principal Component 2") +
+   theme_minimal()
```



Conclusion:

PCA successfully reduced the dimensionality of the air travel dataset while retaining most of the variance. The first two principal components reveal clear seasonal patterns, with similar months clustering together. This confirms that PCA is effective for identifying trends in time-series air travel data.

For Faculty Use

Correction Parameters	**Formative Assessment [40%]**	**Timely completion of Practical [40%]**	**Attendance / Learning Attitude [20%]**	
Marks Obtained				