Final Project

suneeth kunche

2025-03-28

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
#install.packages("ggplot2")
#install.packages("dplyr")
#install.packages("factoextra")
#install.packages("glmnet")
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

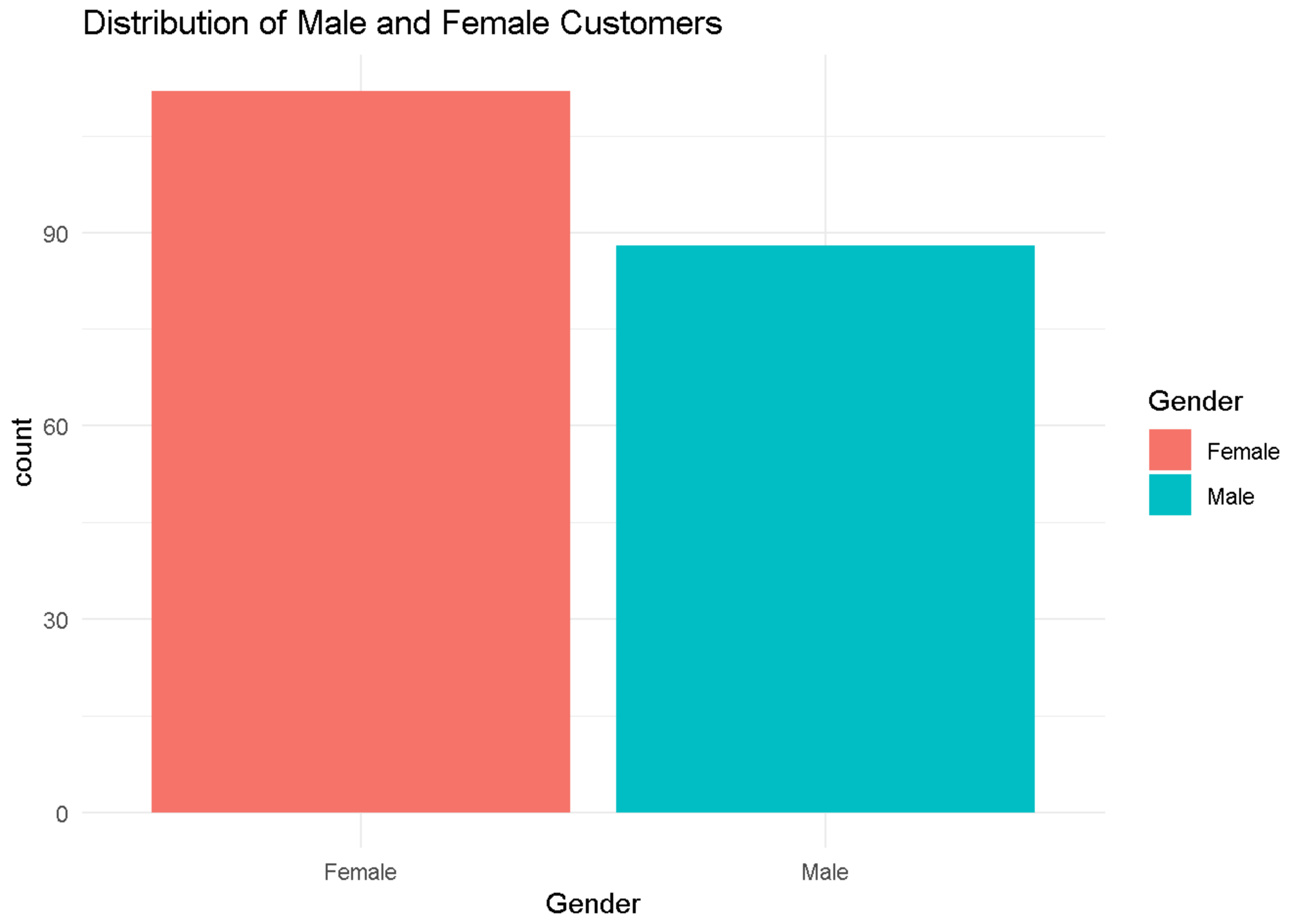
```
file_path <- "C:/Users/Asus/OneDrive/Desktop/dataset/Mall_Customers.csv"


mall_customers <- read.csv(file_path, header = TRUE, stringsAsFactors = FALSE)


head(mall_customers)
```

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                     39
## 2          2   Male  21                 15                     81
## 3          3 Female  20                 16                      6
## 4          4 Female  23                 16                     77
## 5          5 Female  31                 17                     40
## 6          6 Female  22                 17                     76
```

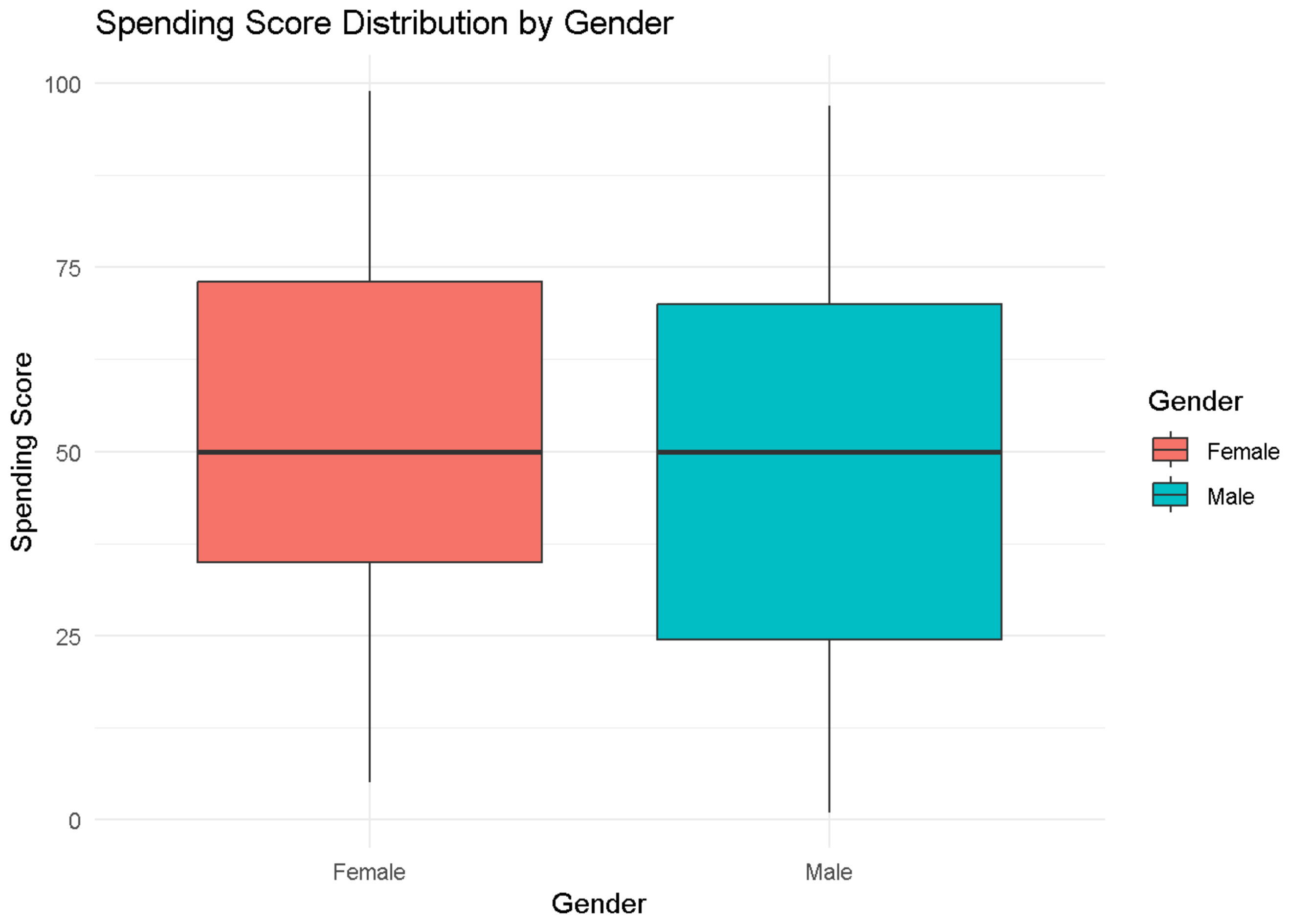
```
colnames(mall_customers)
```

```
## [1] "CustomerID"             "Gender"                 "Age"
## [4] "Annual.Income..k.."     "Spending.Score..1.100."
```

```
ggplot(mall_customers, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  labs(title = "Distribution of Male and Female Customers", x = "Gender", y = "count") +
  theme_minimal()
```



```
ggplot(mall_customers, aes(x = Gender, y = Spending.Score..1.100., fill = Gender)) +
  geom_boxplot() +
  labs(title = "Spending Score Distribution by Gender", x = "Gender", y = "Spending Score") +
  theme_minimal()
```


Spending Score Distribution by Gender

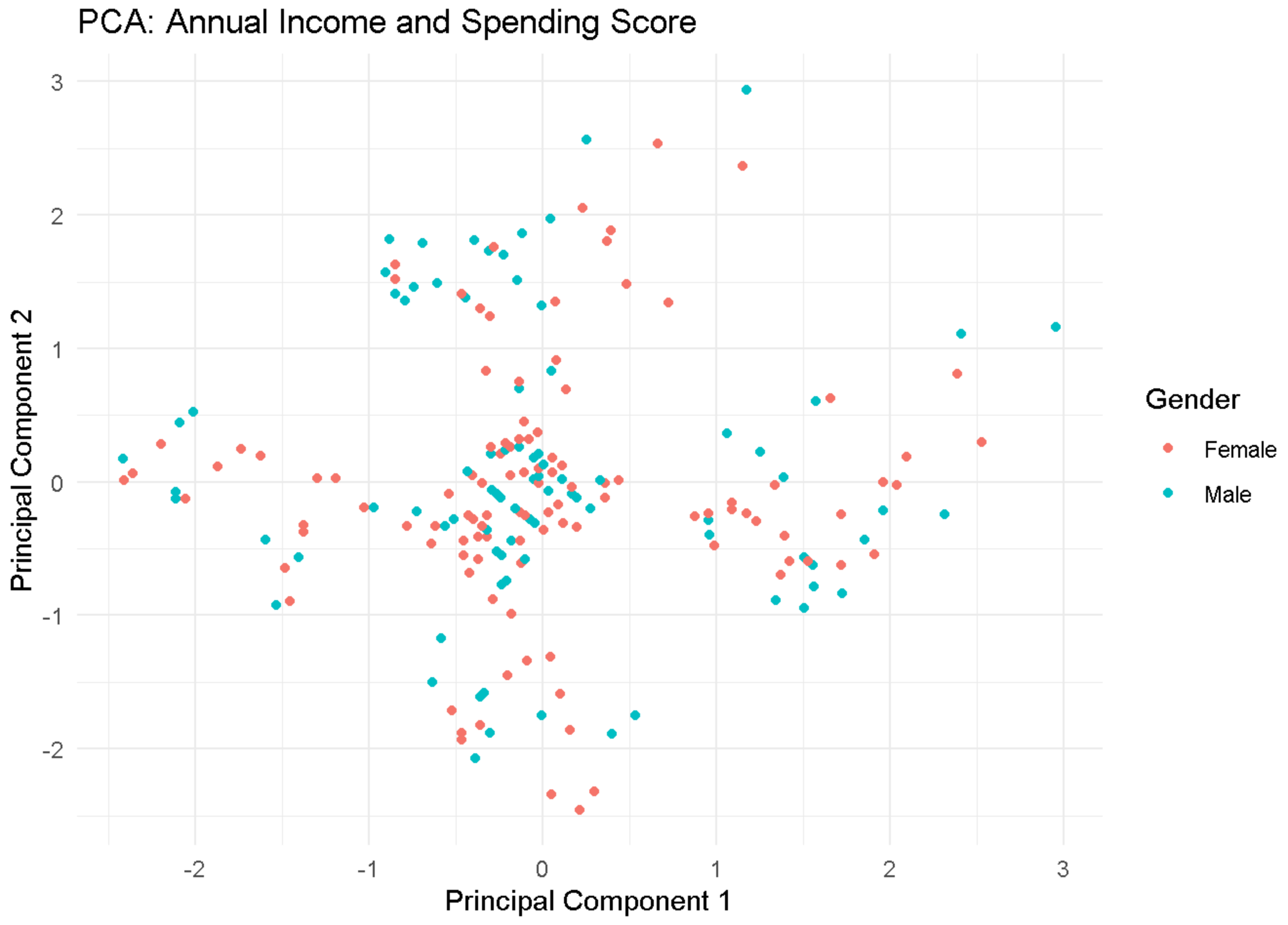
```
# Load required libraries
library(ggplot2)
library(dplyr)

# Select numeric features: Annual Income, Spending Score
selected_data <- mall_customers %>%
  select(`Annual.Income..k..`, `Spending.Score..1.100.`)

# Perform PCA on selected features
pca <- prcomp(selected_data, center = TRUE, scale. = TRUE)


summary(pca)
```

```
## Importance of components:
##                          PC1   PC2
## Standard deviation     1.005 0.995
## Proportion of Variance 0.505 0.495
## Cumulative Proportion  0.505 1.000
```

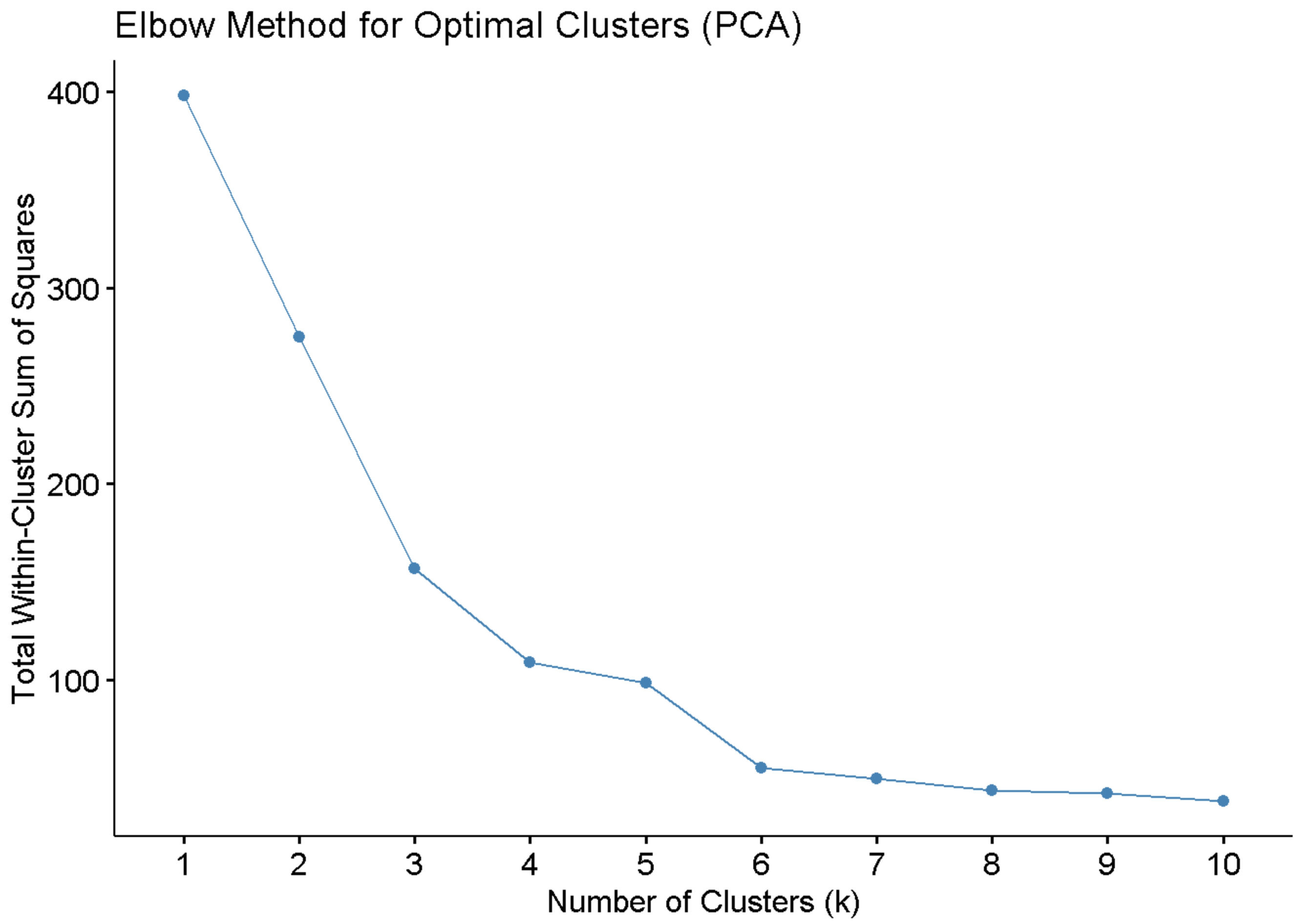
```
pca_data <- as.data.frame(pca$x)


pca_data$Gender <- mall_customers$Gender

# View the new dataset
head(pca_data)
```

```
##          PC1         PC2 Gender
## 1 -1.5332616 -0.91989864   Male
## 2 -0.3832060 -2.06995421   Male
## 3 -2.4099544  0.01063875 Female
## 4 -0.4658128 -1.93350280 Female
## 5 -1.4520347 -0.89343631 Female
## 6 -0.4662728 -1.87919822 Female
```

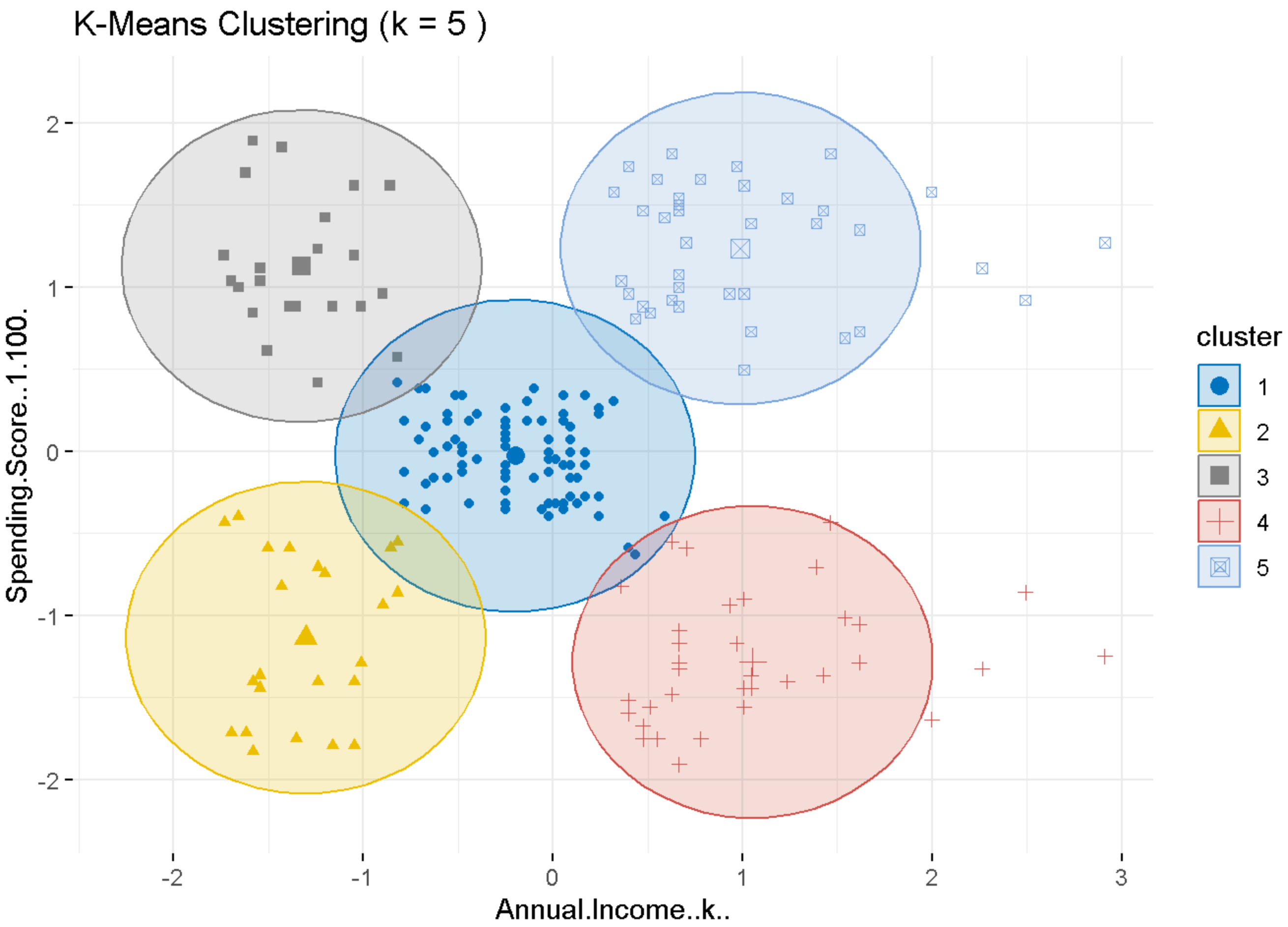
```
# Visualize the first two principal components, colored by Gender
ggplot(pca_data, aes(x = PC1, y = PC2, color = Gender)) +
  geom_point() +
  labs(title = "PCA: Annual Income and Spending Score",
       x = "Principal Component 1",
       y = "Principal Component 2") +
  theme_minimal()
```



```
# Load required libraries
library(factoextra)
library(dplyr)

# Select the relevant features
clustering_data <- mall_customers %>%
  select(Annual.Income..k.., Spending.Score..1.100.)

# Scale the data
scaled_data <- scale(clustering_data)

# Perform PCA
pca <- prcomp(scaled_data, center = TRUE, scale. = TRUE)

# Create PCA data (scores for the first two principal components)
pca_data <- as.data.frame(pca$x)

# Elbow method to determine optimal number of clusters on PCA data
fviz_nbclust(pca_data, kmeans, method = "wss") +
  labs(title = "Elbow Method for Optimal Clusters (PCA)",
       x = "Number of Clusters (k)",
       y = "Total Within-Cluster Sum of Squares")
```

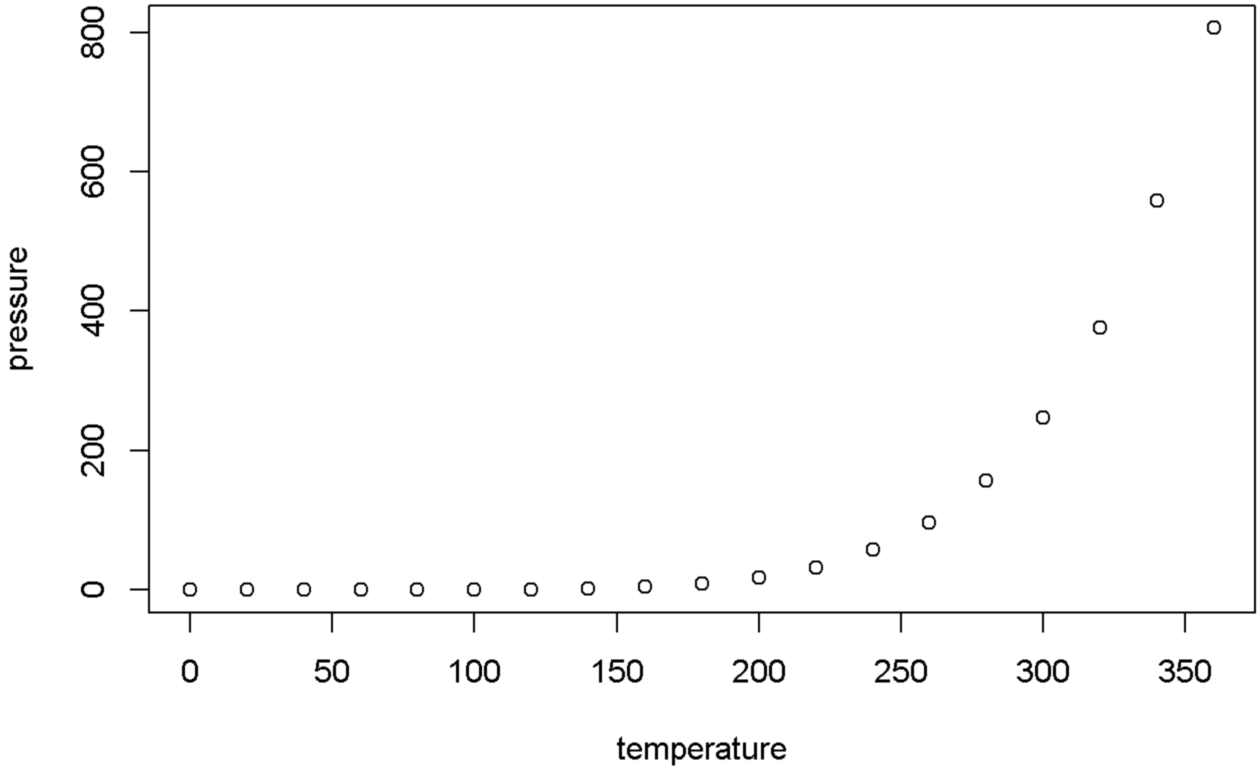


```
set.seed(123)
k <- 5
kmeans_result <- kmeans(pca_data, centers = k, nstart = 25)

# Visualize the clusters
fviz_cluster(kmeans_result, data = scaled_data,
             geom = "point",
             ellipse.type = "euclid",
             palette = "jco",
             ggtheme = theme_minimal()) +
  labs(title = paste("K-Means Clustering (k =", k, ")"))
```



Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.