K-means

2023-11-16

```
options(repos = "https://cran.r-project.org")
```

Load and Explore the Data

```
# Load the dataset

pharma_data <- read.csv("Pharmaceuticals.csv")
```

```
# structure of the dataset

str(pharma_data)
```

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZeneca PL
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

```
# summary of the dataset

summary(pharma_data)
```

```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio           ROE             ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
```

```
##  Min.   :-3.17   Min.   : 2.6      Length:21            Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character     Class :character
##  Median : 9.37   Median :16.1      Mode  :character     Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

```
# First few rows of the data

head(pharma_data)
```

```
##   Symbol                Name Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover
## 1    ABT Abbott Laboratories      68.44 0.32     24.7 26.4 11.8            0.7
## 2    AGN      Allergan, Inc.       7.58 0.41     82.5 12.9  5.5            0.9
## 3    AHM        Amersham plc       6.30 0.46     20.7 14.9  7.8            0.9
## 4    AZN     AstraZeneca PLC      67.63 0.52     21.5 27.4 15.4            0.9
## 5    AVE             Aventis      47.16 0.32     20.1 21.8  7.5            0.6
## 6    BAY            Bayer AG      16.90 1.11     27.9  3.9  1.4            0.6
##   Leverage Rev_Growth Net_Profit_Margin Median_Recommendation Location Exchange
## 1     0.42       7.54              16.1          Moderate Buy       US     NYSE
## 2     0.60       9.16               5.5          Moderate Buy   CANADA     NYSE
## 3     0.27       7.05              11.2            Strong Buy       UK     NYSE
## 4     0.00      15.00              18.0         Moderate Sell       UK     NYSE
## 5     0.34      26.81              12.9          Moderate Buy   FRANCE     NYSE
## 6     0.00      -3.17               2.6                  Hold  GERMANY     NYSE
```

Data Preprocessing

Checking for Missing Values

```
missing_values <- colSums(is.na(pharma_data))

print(missing_values[missing_values > 0])
```

```
## named numeric(0)
```

Handling Missing Values

```
pharma_data_complete <- na.omit(pharma_data)
```

Feature Selection and Scaling

```
numeric_columns <- pharma_data[, sapply(pharma_data, is.numeric)]

numeric_columns[is.na(numeric_columns)] <- apply(numeric_columns, 2, function(x) mean(x, na.rm = TRUE))

scaled_data <- scale(numeric_columns)
```

k-means clustering

```
set.seed(123)
k <- 3
```

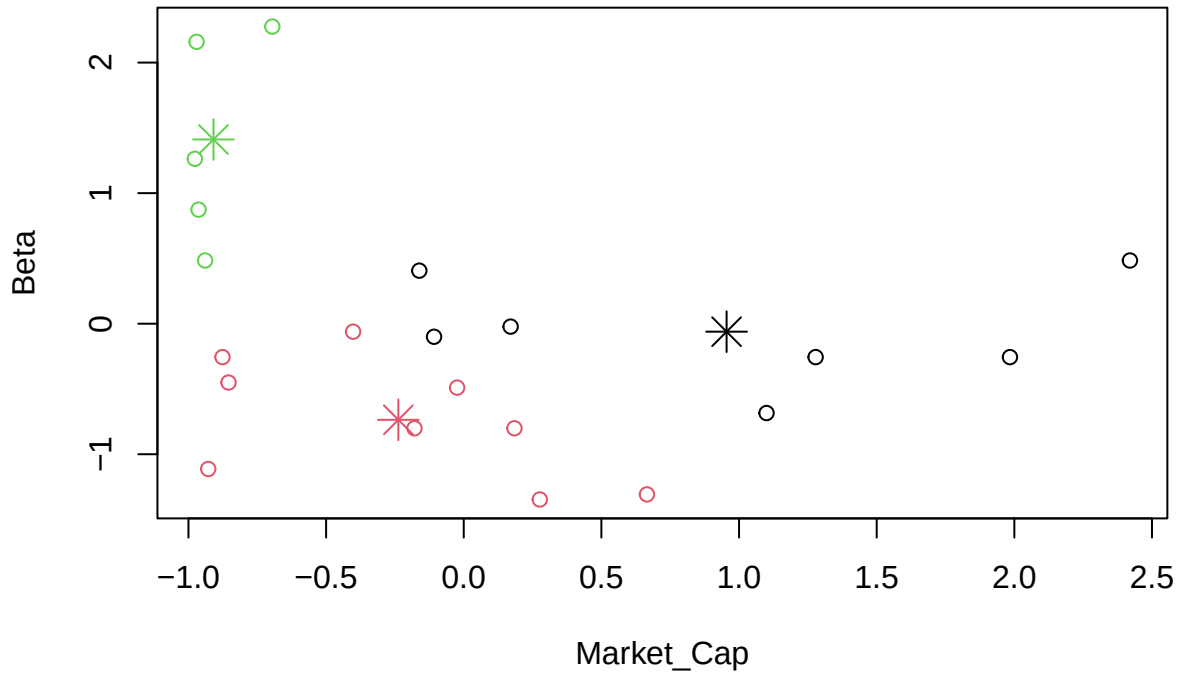
```
kmeans_model <- kmeans(scaled_data, centers = k)

kmeans_model$cluster
```

```
##  [1] 2 2 2 1 2 3 1 3 3 2 1 3 1 3 1 2 1 2 2 2 1
```

```
kmeans_model$centers
```

```
##   Market_Cap        Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.9547543 -0.06120687 -0.3576482  1.0818081  1.1033619      0.8566361
## 2 -0.2375550 -0.73633718  0.4233386 -0.4489909 -0.2407172     -0.1025035
## 3 -0.9090570  1.41109654 -0.2613021 -0.7063477 -1.1114156     -1.0147843
##     Leverage  Rev_Growth Net_Profit_Margin
## 1 -0.2797499 -0.01818848         0.7082574
## 2 -0.3557313 -0.13595383        -0.1652117
## 3  1.0319661  0.27018076        -0.6941793
```

```
plot(scaled_data, col = kmeans_model$cluster)
points(kmeans_model$centers, col = 1:k, pch = 8, cex = 2)
```



Interpretation of Clusters

```
cluster_assignments <- kmeans_model$cluster

scaled_data_df <- as.data.frame(scaled_data)

non_numeric_names <- names(pharma_data)[-c(1:9)]
renamed_pharma_data <- pharma_data
```

```
names(renamed_pharma_data)[which(names(renamed_pharma_data) %in% non_numeric_names)] <- paste0(non_numer

renamed_pharma_data_numeric <- renamed_pharma_data[, sapply(renamed_pharma_data, is.numeric)]

clustered_data <- cbind(scaled_data_df, Cluster = cluster_assignments)
clustered_data <- cbind(clustered_data, renamed_pharma_data_numeric)

dup_cols <- names(clustered_data)[duplicated(names(clustered_data))]
clustered_data <- setNames(clustered_data, make.unique(names(clustered_data), sep = "_"))

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
```

```
cluster_summary <- clustered_data %>%
  group_by(Cluster) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE), .groups = 'drop')
```

```
## Warning: There was 1 warning in `summarise()`.
## i In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.
## i In group 1: `Cluster = 1`.
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

```
print(cluster_summary)
```

```
## # A tibble: 3 x 19
##   Cluster Market_Cap    Beta PE_Ratio    ROE    ROA Asset_Turnover Leverage
##     <int>      <dbl>   <dbl>    <dbl>  <dbl>  <dbl>          <dbl>    <dbl>
## 1       1      0.955 -0.0612   -0.358  1.08   1.10           0.857   -0.280
## 2       2     -0.238 -0.736     0.423 -0.449 -0.241         -0.103   -0.356
## 3       3     -0.909  1.41     -0.261 -0.706 -1.11          -1.01     1.03
## # i 11 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Market_Cap_1 <dbl>, Beta_1 <dbl>, PE_Ratio_1 <dbl>, ROE_1 <dbl>,
## #   ROA_1 <dbl>, Asset_Turnover_1 <dbl>, Leverage_1 <dbl>,
## #   Rev_Growth_orig <dbl>, Net_Profit_Margin_orig <dbl>
```

Analyze Other Variables

```
unique(pharma_data$Median_Recommendation)
```

```
## [1] "Moderate Buy"  "Strong Buy"    "Moderate Sell" "Hold"
```

```
sum(is.na(pharma_data$Median_Recommendation))
```

```
## [1] 0
```

```
sum(pharma_data$Median_Recommendation == "")
```

```
## [1] 0
```

```
length(pharma_data$Median_Recommendation)
```

```
## [1] 21
```

```
colnames(pharma_data)
```

```
##  [1] "Symbol"                "Name"                  "Market_Cap"
##  [4] "Beta"                  "PE_Ratio"              "ROE"
##  [7] "ROA"                   "Asset_Turnover"        "Leverage"
## [10] "Rev_Growth"            "Net_Profit_Margin"     "Median_Recommendation"
## [13] "Location"              "Exchange"
```

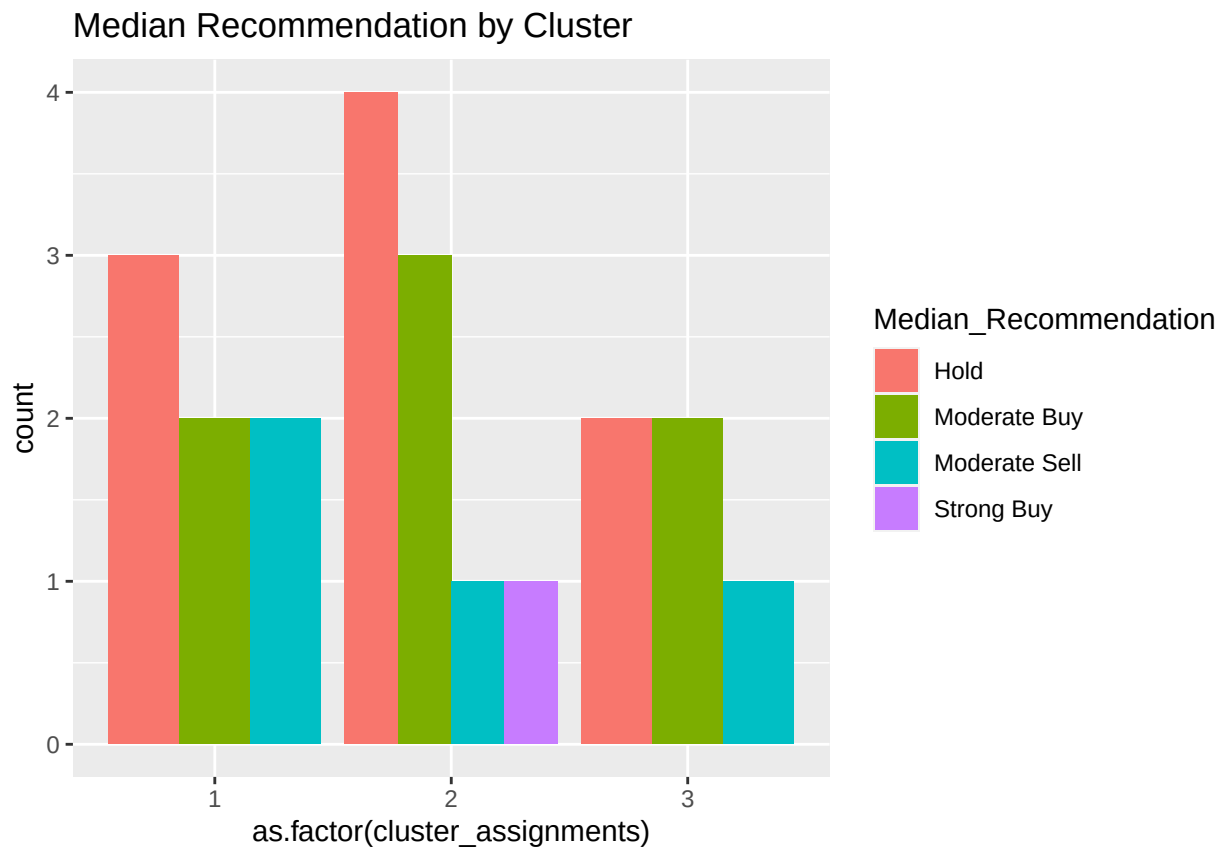
```
library(ggplot2)

table_median_recommendation <- table(cluster_assignments, pharma_data$Median_Recommendation)

print("Table for Median Recommendation:")
```

```
## [1] "Table for Median Recommendation:"
```

```
print(table_median_recommendation)
```

```
##
## cluster_assignments Hold Moderate Buy Moderate Sell Strong Buy
##                   1    3            2             2          0
##                   2    4            3             1          1
##                   3    2            2             1          0
```

```
ggplot(data = pharma_data, aes(x = as.factor(cluster_assignments), fill = Median_Recommendation)) +
  geom_bar(position = "dodge") +
  labs(title = "Median Recommendation by Cluster")
```



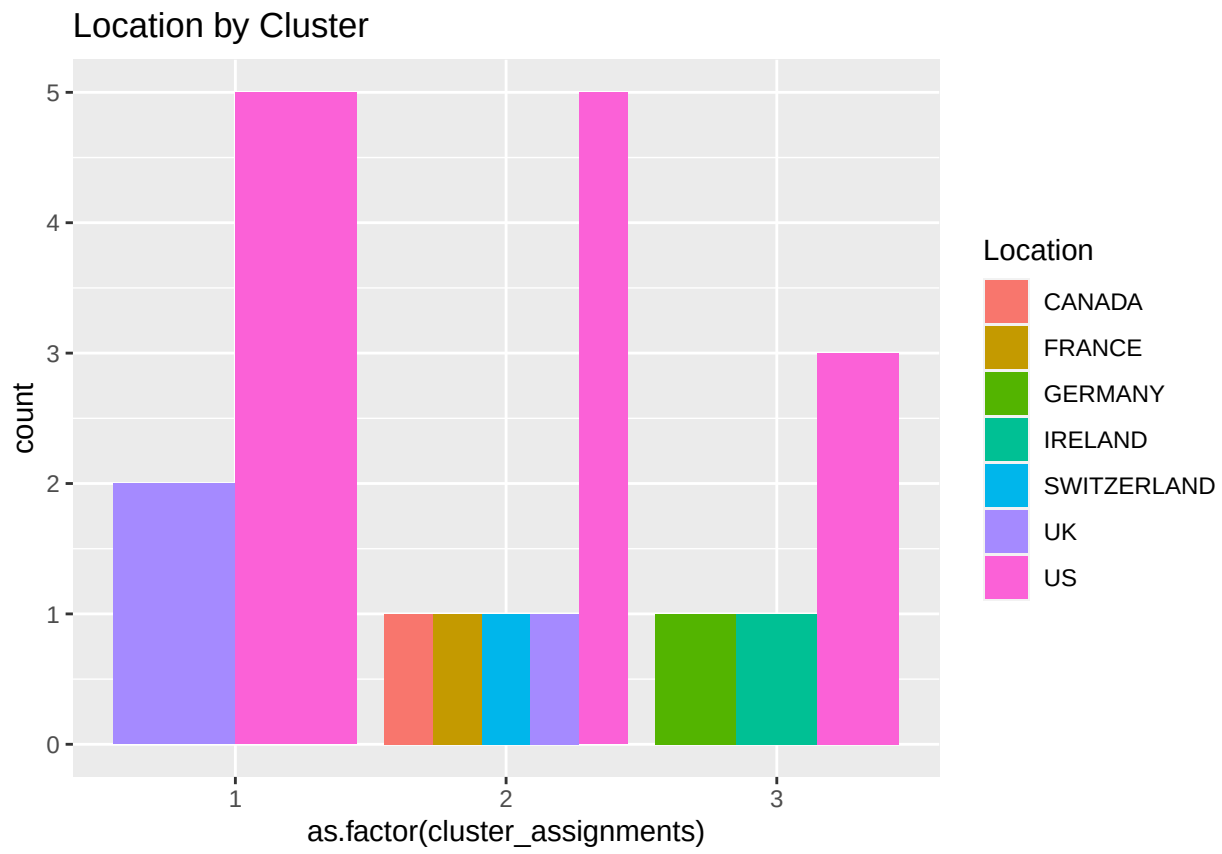

```
table_location <- table(cluster_assignments, pharma_data$Location)

print("Table for Headquarters:")
```

```
## [1] "Table for Headquarters:"
```

```
print(table_location)
```

```
##
## cluster_assignments CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US
##                   1      0      0       0       0           0  2  5
##                   2      1      1       0       0           1  1  5
##                   3      0      0       1       1           0  0  3
```

```
ggplot(data = pharma_data, aes(x = as.factor(cluster_assignments), fill = Location)) +
  geom_bar(position = "dodge") +
  labs(title = "Location by Cluster")
```



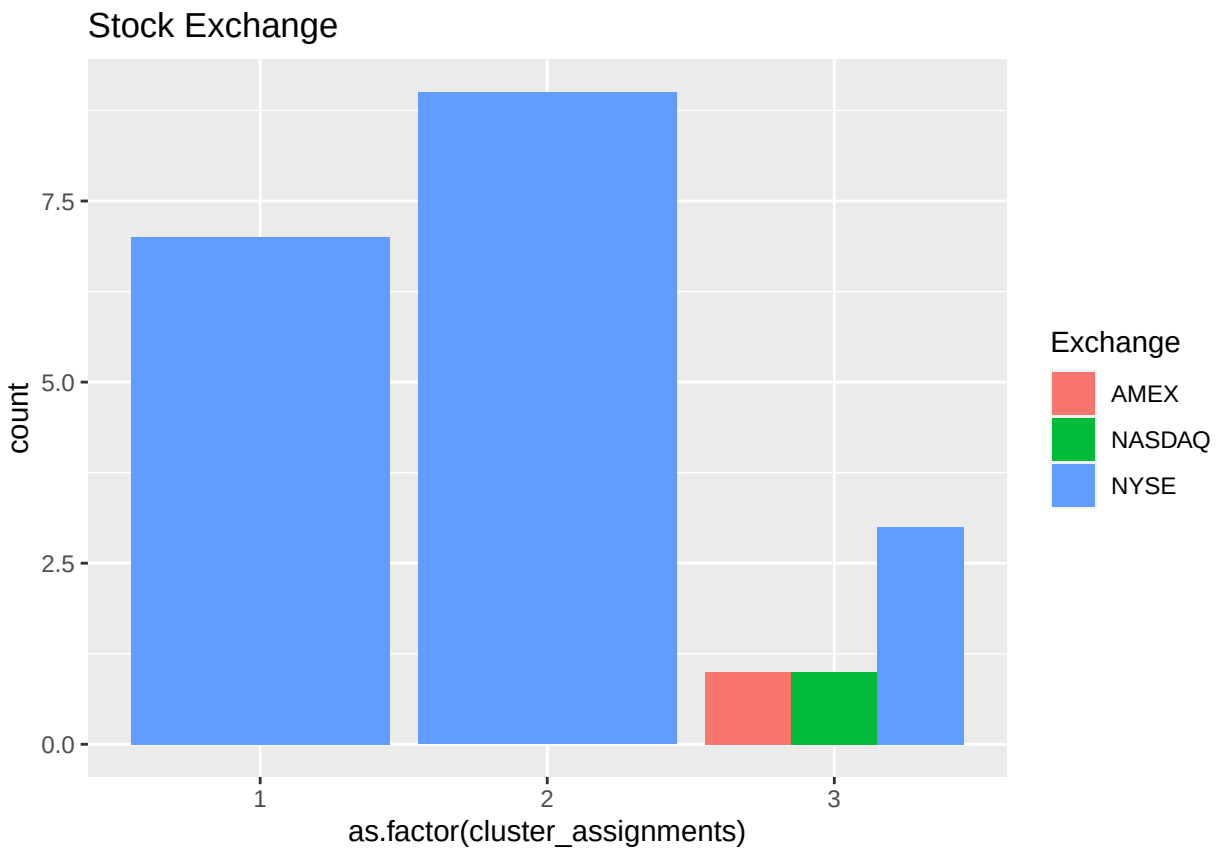

```
table_stock_exchange <- table(cluster_assignments, pharma_data$Exchange)

print("Table for Stock Exchange:")
```

```
## [1] "Table for Stock Exchange:"
```

```
print(table_stock_exchange)
```

```
##
## cluster_assignments AMEX NASDAQ NYSE
##                   1    0      0    7
##                   2    0      0    9
##                   3    1      1    3
```

```
ggplot(data = pharma_data, aes(x = as.factor(cluster_assignments), fill = Exchange)) +
  geom_bar(position = "dodge") +
  labs(title = "Stock Exchange")
```



Cluster Names

```
num_clusters <- 3

cluster_centroids <- kmeans_model$centers

assign_cluster_names <- function(centroids) {
  cluster_names <- character(num_clusters)
  for (i in 1:num_clusters) {
     if (centroids[i, "Market_Cap"] > 50) {
       cluster_names[i] <- "High Market Cap"
    } else if (centroids[i, "Market_Cap"] < 20) {
       cluster_names[i] <- "Low Market Cap"
    } else {
       cluster_names[i] <- "Moderate Market Cap"
    }
  }
  return(cluster_names)
}


cluster_names <- assign_cluster_names(cluster_centroids)

cluster_names
```

```
## [1] "Low Market Cap" "Low Market Cap" "Low Market Cap"
```