Assignment_4

Alexis McCartney

October 25, 2025

Contents

About
This report applies k-means clustering to data for 21 pharmaceutical firms using nine financial measures. Variables are standardized to equalize scale. The number of clusters (k) is chosen using elbow and silhouette diagnostics.
Cluster profiles are interpreted and patterns in variables not used for clustering are examined.

0.1 (A) Clustering with Numerical Variables (1–9)

Method & justification.
Nine financial measures (Market Cap, Beta, P/E, ROE, ROA, Asset Turnover, Leverage, Revenue Growth, Net Profit Margin) are standardized so each has equal weight.
K-means is run with multiple starts (`nstart = 50`) and k is chosen via elbow & silhouette diagnostics.

0.1.1 1) Import data

```
Pharmaceuticals <- read.csv("Pharmaceuticals.csv", header=TRUE, stringsAsFactors=FALSE)
stopifnot(is.data.frame(Pharmaceuticals))
```

0.1.2 2) Select numeric features

```
num_cols <- c("Market_Cap","Beta","PE_Ratio","ROE","ROA",
              "Asset_Turnover","Leverage","Rev_Growth","Net_Profit_Margin")
Pharma_numeric <- Pharmaceuticals[, num_cols]
```

0.1.3 3) Coerce to numeric (clean commas, % etc.)

```r
Pharma_numeric <- as.data.frame(lapply(Pharma_numeric, function(x){
  x <- gsub(",", "", as.character(x))
  x <- gsub("%", "", x)
  x <- trimws(x)
  suppressWarnings(as.numeric(x))
}))
stopifnot(all(sapply(Pharma_numeric, is.numeric)))
```

0.1.4 4) Median-impute NAs

```r
for(j in seq_len(ncol(Pharma_numeric))){
  if(anyNA(Pharma_numeric[[j]])){
    Pharma_numeric[[j]][is.na(Pharma_numeric[[j]])] <- median(Pharma_numeric[[j]], na.rm=TRUE)
  }
}
```

0.1.5 5) Drop zero/undefined-variance columns

```r
sds <- sapply(Pharma_numeric, sd)
drop_cols <- names(sds)[!is.finite(sds) | sds==0]
if(length(drop_cols)>0){
  message("Dropping: ", paste(drop_cols, collapse=", "))
  Pharma_numeric <- Pharma_numeric[, setdiff(names(Pharma_numeric), drop_cols), drop=FALSE]
}
```

0.1.6 6) Standardize (z-scores)

```r
Pharma_scaled <- scale(Pharma_numeric)
stopifnot(all(is.finite(as.matrix(Pharma_scaled))))
set.seed(123)
```

0.1.7 Elbow plot

```r
wss <- sapply(1:8, function(k){
  kmeans(Pharma_scaled, centers=k, nstart=50, iter.max=100)$tot.withinss
})
plot(1:8, wss, type="b", xlab="k", ylab="Total WSS", main="Elbow Plot")
```

0.1.8 Silhouette plot

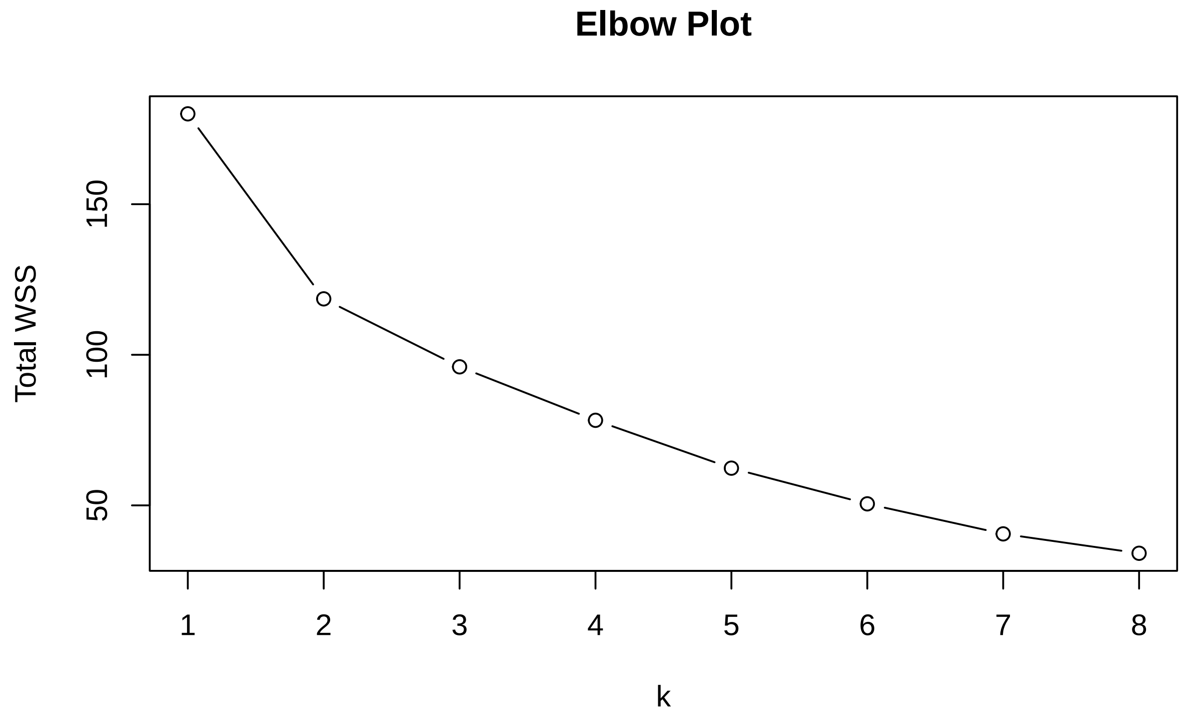

Figure 1: Elbow Plot

```
sil_mean <- sapply(2:8, function(k){
  km <- kmeans(Pharma_scaled, centers=k, nstart=50, iter.max=100)
  ss <- silhouette(km$cluster, dist(Pharma_scaled))
  mean(ss[,3])
})
plot(2:8, sil_mean, type="b", xlab="k", ylab="Mean Silhouette", main="Silhouette Plot")
```

0.1.9 7) Choose k

```
K_CHOSEN <- 3
K_CHOSEN
```

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

0.1.10 8) Fit final k-means

```
set.seed(123)
km_final <- kmeans(Pharma_scaled, centers=K_CHOSEN, nstart=50, iter.max=100)
```

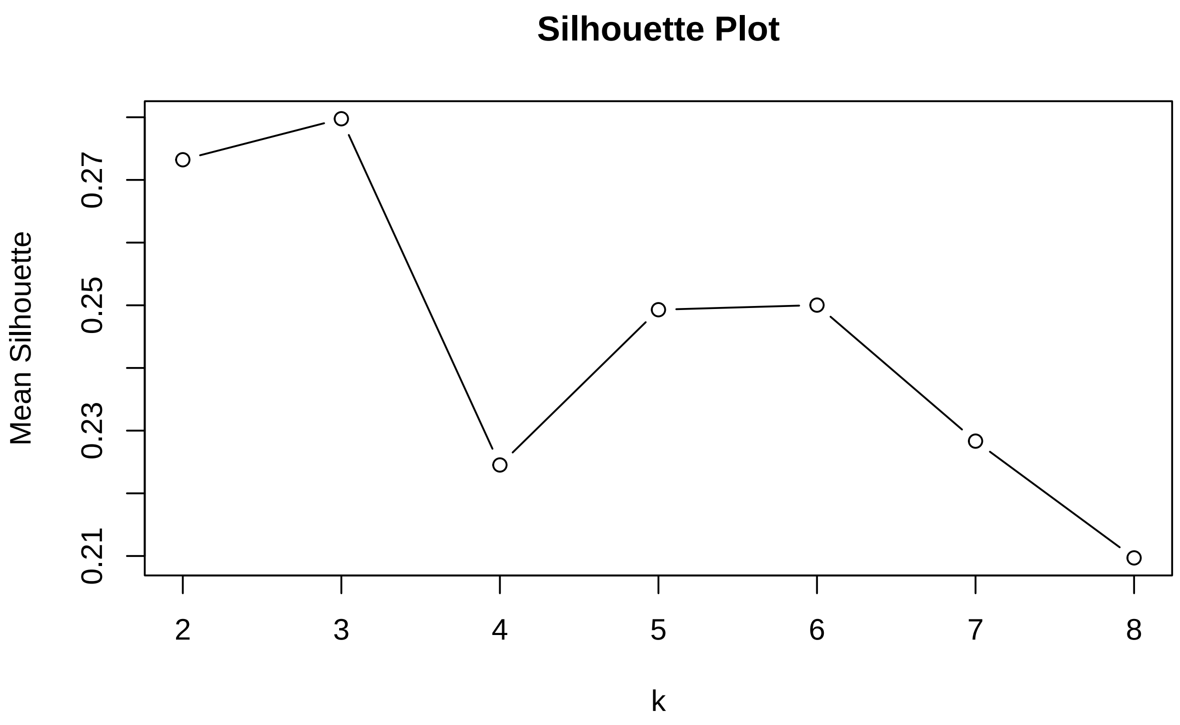

Figure 2: Silhouette Mean by k

0.1.11 Add cluster labels

```
Pharmaceuticals$Cluster <- factor(km_final$cluster)
```

0.1.12 Cluster sizes and centers

```
sizes <- km_final$size
centers <- round(km_final$centers, 2)
kbl(data.frame(Cluster=seq_along(sizes), Size=sizes), caption="Cluster Sizes")
```

Table 1: Cluster Sizes

Cluster	Size
1	4
2	11
3	6

```
kbl(as.data.frame(centers), caption="Standardized Cluster Centers (z-scores)")
```

Table 2: Standardized Cluster Centers (z-scores)

Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin
-0.61	0.27	1.31	-0.96	-1.02	0.23	-0.36	-0.58	-1.38
0.67	-0.36	-0.28	0.66	0.83	0.46	-0.33	-0.29	0.68
-0.83	0.48	-0.37	-0.56	-0.85	-1.00	0.85	0.92	-0.33

0.1.13 Unscaled means by cluster

```
cluster_profile <- aggregate(Pharma_numeric, by=list(Cluster=Pharmaceuticals$Cluster), mean)
cluster_profile_fmt <- cluster_profile
cluster_profile_fmt[-1] <- lapply(cluster_profile_fmt[-1], function(x) round(x,2))
kbl(cluster_profile_fmt, caption="Unscaled Feature Means by Cluster")
```

Table 3: Unscaled Feature Means by Cluster

Cluster	Market_Cap	Beta	PE_Ratio	ROE	ROA	Asset_Turnover	Leverage	Rev_Growth	Net_Profit_Margin
1	21.75	0.60	46.90	11.3	5.10	0.75	0.30	7.01	6.65
2	97.11	0.43	20.95	35.7	14.95	0.80	0.33	10.16	20.17
3	9.23	0.65	19.43	17.3	5.98	0.48	1.25	23.49	13.52

0.2 (B) Interpret the Clusters

Cluster 1 – Large efficient profitable firms (high ROE/ROA, high asset turnover, low leverage).
Cluster 2 – Mid-caps with higher P/E and moderate profitability (priced for growth).
Cluster 3 – Small volatile firms (high beta & leverage; low profitability).

```
centers_df <- as.data.frame(km_final$centers)
centers_df$Cluster <- factor(rownames(centers_df))
var_names <- setdiff(names(centers_df), "Cluster")

centers_long <- data.frame(
  Cluster = rep(centers_df$Cluster, each=length(var_names)),
  Variable = rep(var_names, times=nrow(centers_df)),
  z = unlist(centers_df[var_names], use.names=FALSE)
)

plot(centers_long$z ~ interaction(centers_long$Cluster, centers_long$Variable),
     xlab="Cluster.Variable", ylab="Standardized Mean (z)",
     main="Cluster Profiles (Standardized)", pch=16, las=2, cex.axis=0.7)
abline(h=0, lty=2)
```

0.3 (C) Variables Not Used for Clustering (10–12)

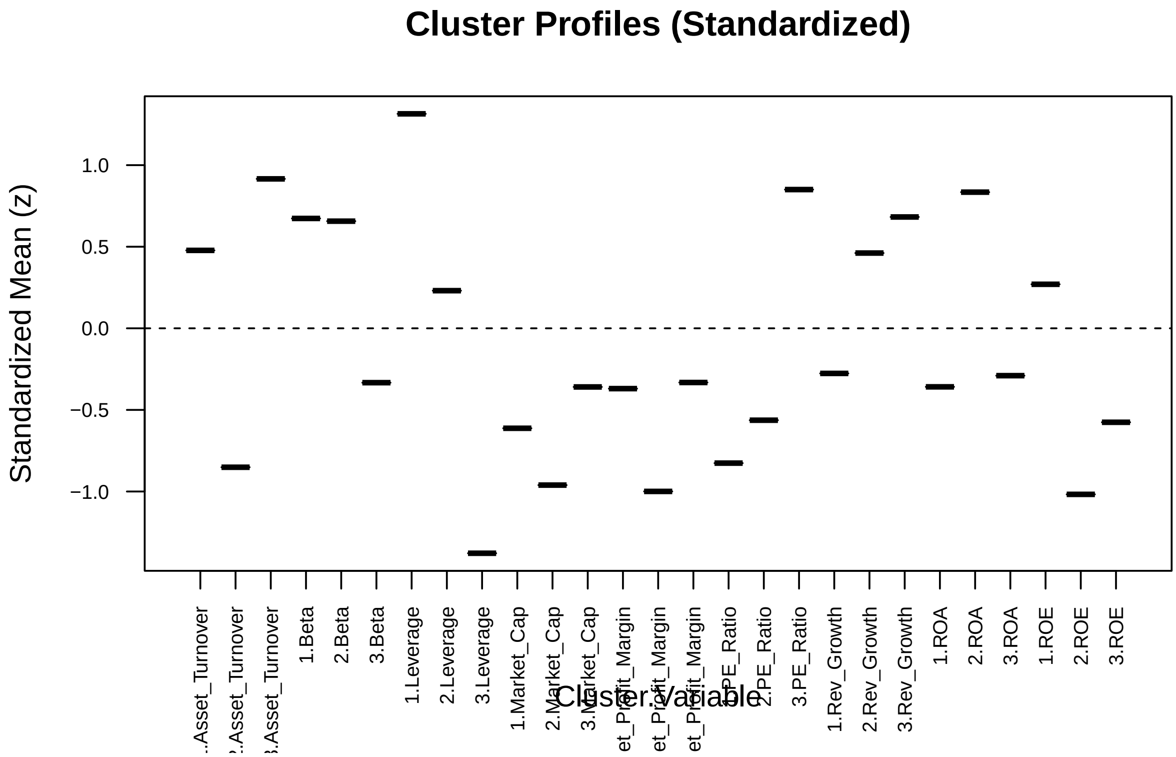

Figure 3: Cluster Profiles (Standardized)

```
for(j in 10:11){
  x <- Pharmaceuticals[[j]]
  x <- gsub(",", "", as.character(x))
  x <- gsub("%", "", x)
  Pharmaceuticals[[j]] <- suppressWarnings(as.numeric(x))
}
Pharmaceuticals$Median_Recommendation <- as.factor(Pharmaceuticals$Median_Recommendation)
```

0.3.1 Numeric means by cluster

```
num_means <- aggregate(Pharmaceuticals[,10:11],
                       by=list(Cluster=Pharmaceuticals$Cluster),
                       mean, na.rm=TRUE)
num_means_fmt <- num_means
num_means_fmt[-1] <- lapply(num_means_fmt[-1], function(x) round(x,2))
kbl(num_means_fmt, caption="Numeric Variables (10-11): Means by Cluster")
```

Table 4: Numeric Variables (10–11): Means by Cluster

Cluster	Rev_Growth	Net_Profit_Margin
1	7.01	6.65

Cluster	Rev_Growth	Net_Profit_Margin
2	10.16	20.17
3	23.49	13.52

0.3.2 Categorical distribution by cluster

```
tab_reco <- table(Pharmaceuticals$Cluster, Pharmaceuticals$Median_Recommendation)
prop_reco <- round(prop.table(tab_reco,1),2)
kbl(as.data.frame.matrix(tab_reco), caption="Counts of Median_Recommendation by Cluster")
```

Table 5: Counts of Median_Recommendation by Cluster

Hold	Moderate Buy	Moderate Sell	Strong Buy
2	1	0	1
6	3	2	0
1	3	2	0

```
kbl(as.data.frame.matrix(prop_reco), caption="Row Proportions of Median_Recommendation by Cluster")
```

Table 6: Row Proportions of Median_Recommendation by Cluster

Hold	Moderate Buy	Moderate Sell	Strong Buy
0.50	0.25	0.00	0.25
0.55	0.27	0.18	0.00
0.17	0.50	0.33	0.00

0.4 (D) Cluster Names

```
cluster_names <- c(
  "1"="Profitable Large-Cap",
  "2"="Valuation-Growth Mid-Cap",
  "3"="Levered Small-Cap"
)
Pharmaceuticals$Cluster_Name <- cluster_names[as.character(Pharmaceuticals$Cluster)]
```

0.4.1 Firms by Named Cluster

```
split(Pharmaceuticals$Name, Pharmaceuticals$Cluster_Name) |> lapply(sort) |> str()
```

```
## List of 3
##  $ Levered Small-Cap       : chr [1:6] "Aventis" "Chattem, Inc" "Elan Corporation, plc" "IVAX Corpora
##  $ Profitable Large-Cap    : chr [1:4] "Allergan, Inc." "Amersham plc" "Bayer AG" "Pharmacia Corporat
##  $ Valuation-Growth Mid-Cap: chr [1:11] "Abbott Laboratories" "AstraZeneca PLC" "Bristol-Myers Squibb
```