Stat4601_StatenIsland_Kmeans

Stephanie Cheng

2025-04-09

K-mean with PCA

Reduce dimensions and prepare data for clustering

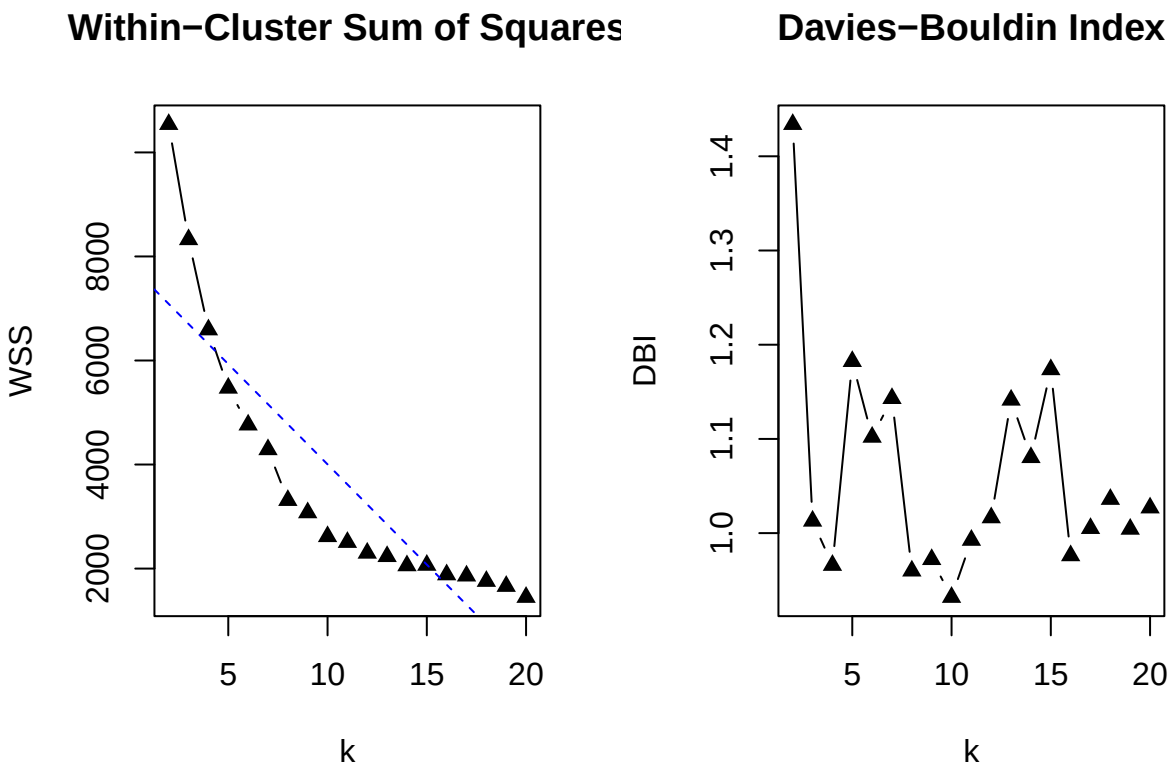
```
SI_pca <- pca(staten_island_data, "Staten Island")
```

```
##
## ===== PCA Summary for Staten Island =====
## Importance of components:
##                             PC1    PC2    PC3     PC4
## Standard deviation     1.3796 1.0253 0.9519 0.37346
## Proportion of Variance 0.4758 0.2628 0.2265 0.03487
## Cumulative Proportion  0.4758 0.7386 0.9651 1.00000
## Contributing variable for each PC:
##                 PC1                 PC2                 PC3                 PC4
##       "TOTAL.UNITS"        "YEAR.BUILT" "GROSS.SQUARE.FEET"       "TOTAL.UNITS"
```

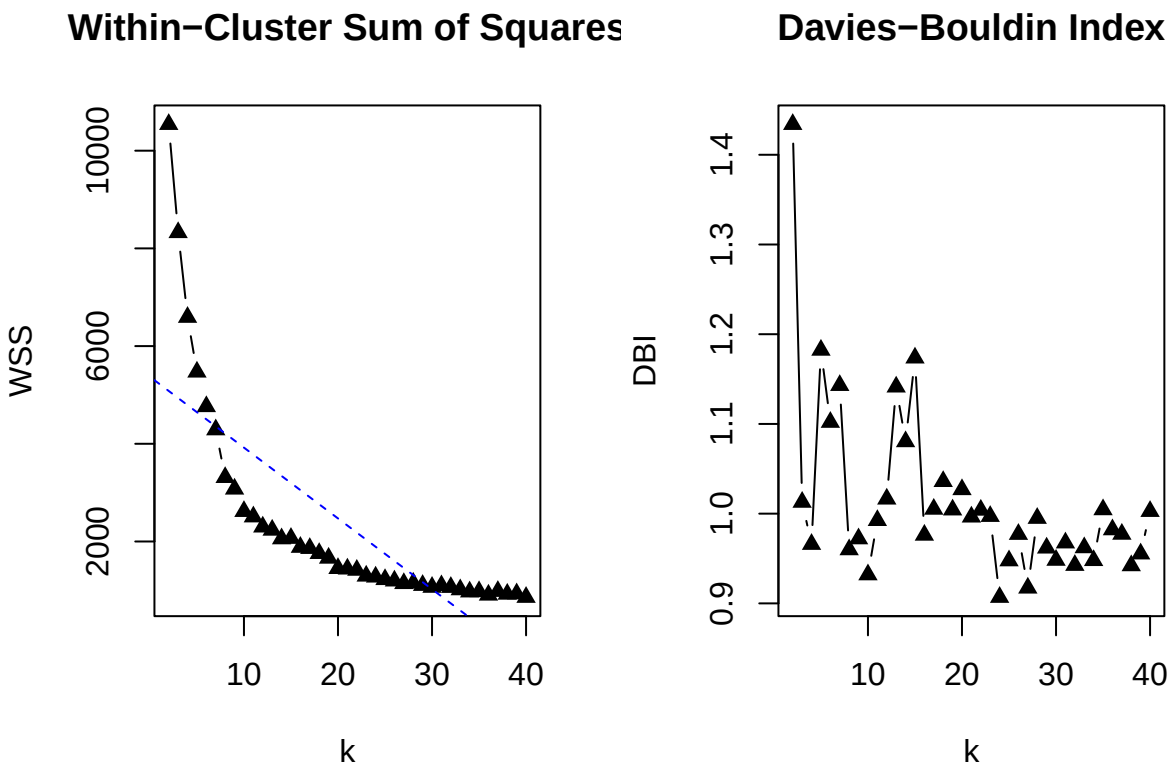
Calculate clustering evaluation with Davies Bouldin index & Within-cluster sum of squares. See the affect when K is increasing, then we can apply elbow method to avoid picking the best k within overfitting case.

```
SI_k_stats_20 <- calculate_k_stats_PCA(SI_pca, max_k = 20)
SI_k_stats_40 <- calculate_k_stats_PCA(SI_pca, max_k = 40)

# DBI & WSS plot
elbows_20 <- plot_kmeans(SI_k_stats_20$errs, SI_k_stats_20$DBI)
```



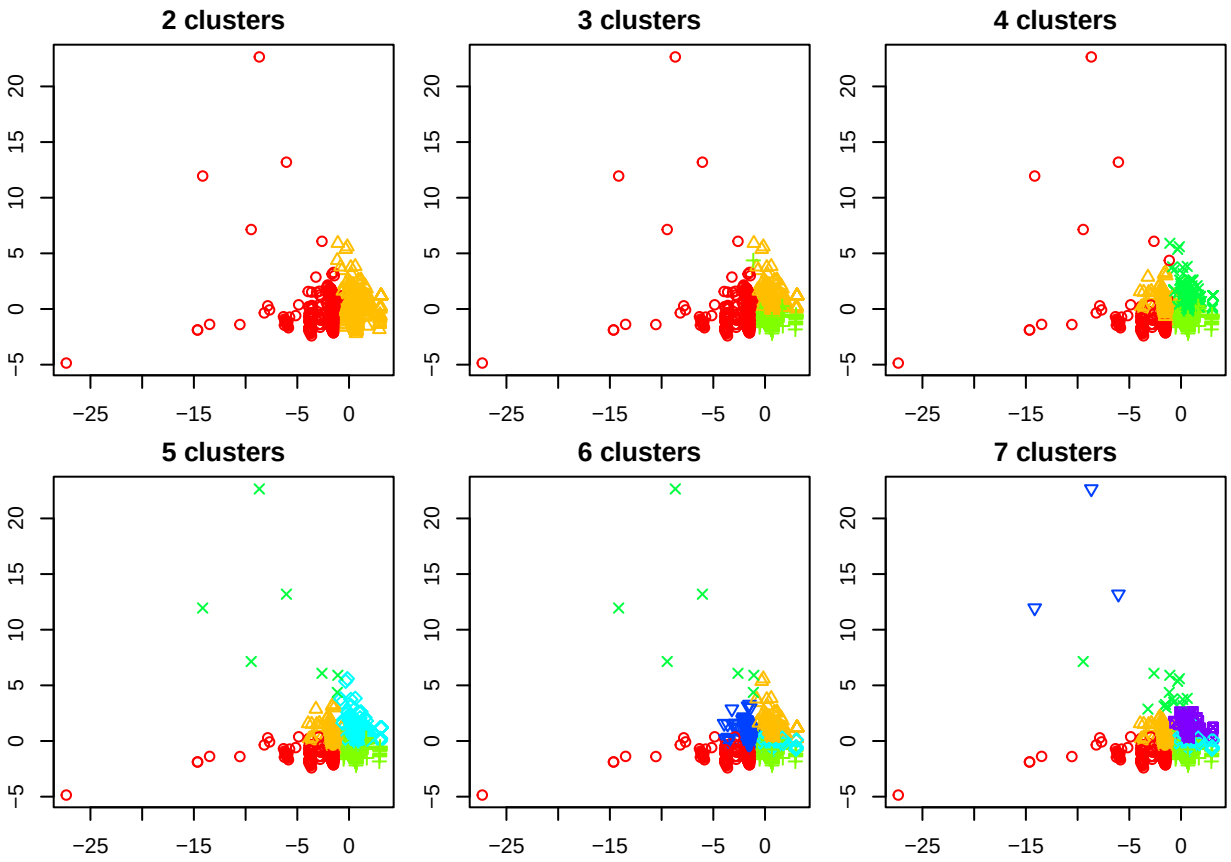

```
elbows_40 <- plot_kmeans(SI_k_stats_40$errs, SI_k_stats_40$DBI)
```




```
best_k <- 4
```

Plot all clusters from 2 to 7 as the best k clusters is within that range.

```
plot_clusters(SI_k_stats_20$X.syn, min_k = 2, max_k = 7)
```



K-means on PCA as PCA gives a lower-dimensional variable that improves clustering quality

```
km <- kmeans(SI_pca$x, centers = best_k, nstart = 25)
summarize_kmeans(km, "Staten Island")
```

```
##
## ===== K-means Model Performance Summary for Staten Island =====
## Total within-cluster sum of squares (WSS): 6582.496
##
## Cluster sizes:
## [1] 1734    5 1099 1065
##
## Cluster centers (in PCA space):
##   TOTAL.UNITS  YEAR.BUILT GROSS.SQUARE.FEET TOTAL.UNITS
## 1   0.7091722  0.54029105        -0.2446298 -0.01910960
## 2  -8.1901592 12.20290706        15.5641803 -1.07366439
## 3   0.6709259 -0.82268801         0.6105251 -0.03589742
## 4  -1.8085459 -0.08802357        -0.3047887  0.07319778
```

Interpret what the clusters mean with the original data

```
staten_island_data$cluster <- km$cluster
aggregate(. ~ cluster, data = staten_island_data, mean)
```

```
##   cluster BOROUGH RESIDENTIAL.UNITS TOTAL.UNITS GROSS.SQUARE.FEET YEAR.BUILT
```

```
## 1       1       5         0.9653979    1.000000          1725.008   1988.572
## 2       2       5         0.0000000    5.000000         41650.600   1970.400
## 3       3       5         0.9599636    1.014559          1439.347   1936.753
## 4       4       5         2.0798122    2.147418          2332.853   1969.888
```

Export the clusters for Supervised learning

```
dir.create("after_cluster_dataset")
```

```
## Warning in dir.create("after_cluster_dataset"): 'after_cluster_dataset' already
## exists
```

```
write.csv(staten_island_data,
          file = "after_cluster_dataset/cleaned_StatenIsland_with_clusters.csv",
          row.names = FALSE)
```