Stat4601_Brooklyn_Kmeans

Stephanie Cheng

2025-04-09

K-mean with PCA

Reduce dimensions and prepare data for clustering

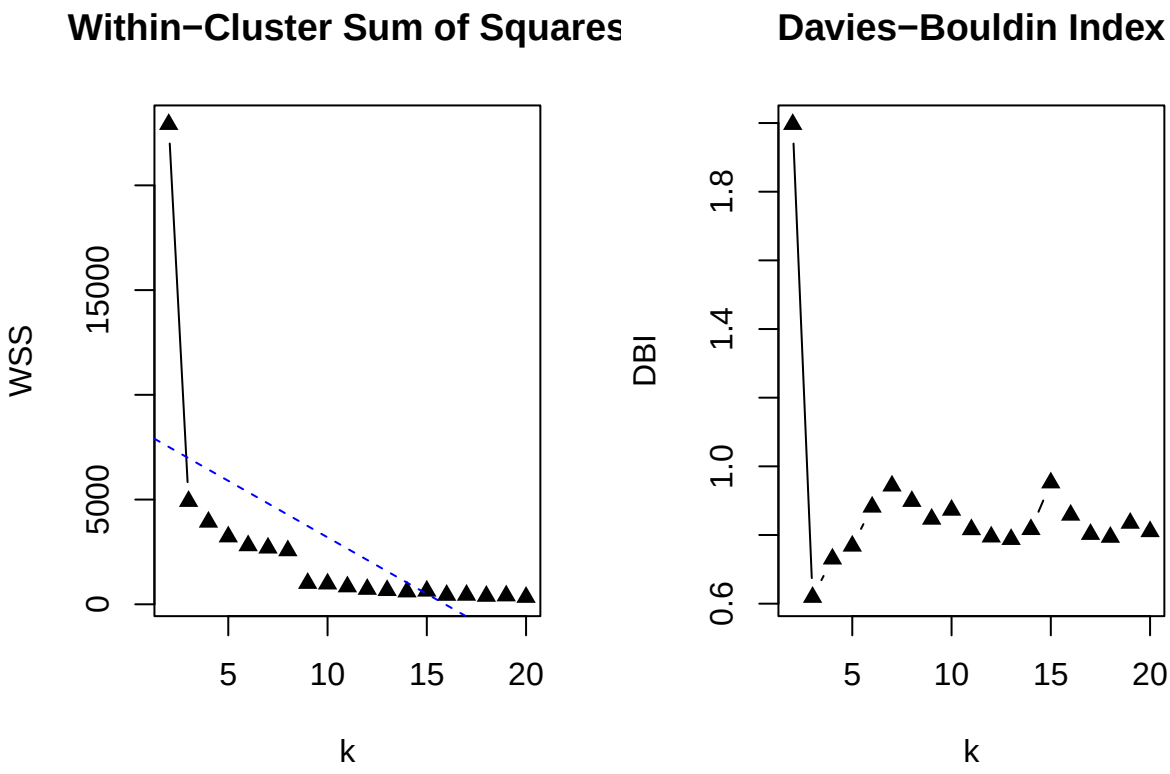
```
brooklyn_pca <- pca(brooklyn_data, "Brooklyn")
```

```
##
## ===== PCA Summary for Brooklyn =====
## Importance of components:
##                            PC1    PC2     PC3     PC4
## Standard deviation     1.739 0.9649 0.20529 0.05174
## Proportion of Variance 0.756 0.2328 0.01054 0.00067
## Cumulative Proportion  0.756 0.9888 0.99933 1.00000
## Contributing variable for each PC:
##                 PC1                 PC2                 PC3                 PC4
##       "TOTAL.UNITS"        "YEAR.BUILT" "GROSS.SQUARE.FEET"       "TOTAL.UNITS"
```

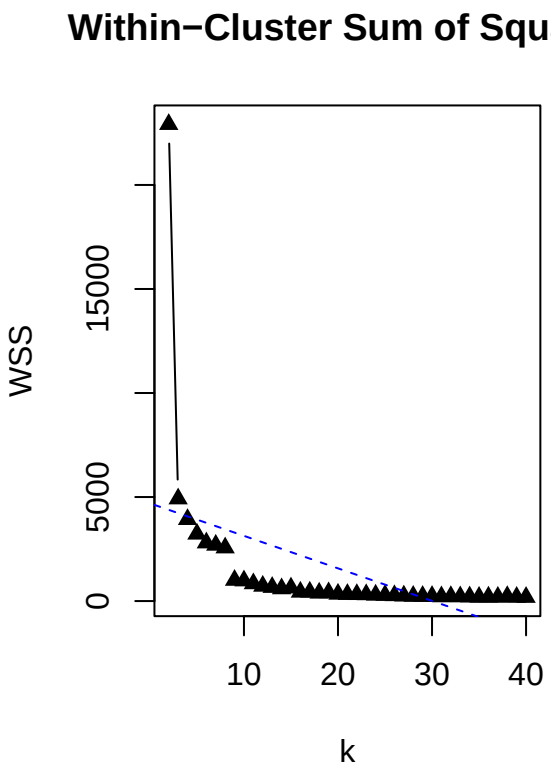
Calculate clustering evaluation with Davies Bouldin index & Within-cluster sum of squares. See the affect when K is increasing, then we can apply elbow method to avoid picking the best k within overfitting case.

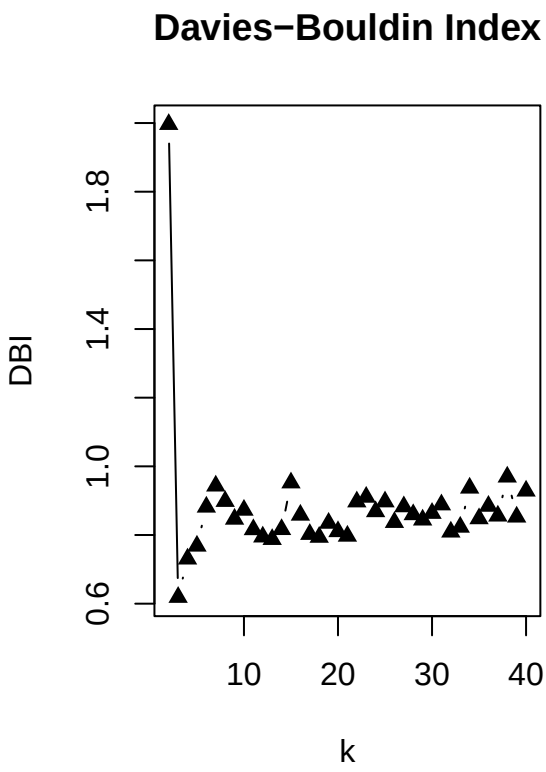
```
brooklyn_k_stats_20 <- calculate_k_stats_PCA(brooklyn_pca, max_k = 20)
brooklyn_k_stats_40 <- calculate_k_stats_PCA(brooklyn_pca, max_k = 40)

# DBI & WSS plot
elbows_20 <- plot_kmeans(brooklyn_k_stats_20$errs, brooklyn_k_stats_20$DBI)
```




```
elbows_40 <- plot_kmeans(brooklyn_k_stats_40$errs, brooklyn_k_stats_40$DBI)
```

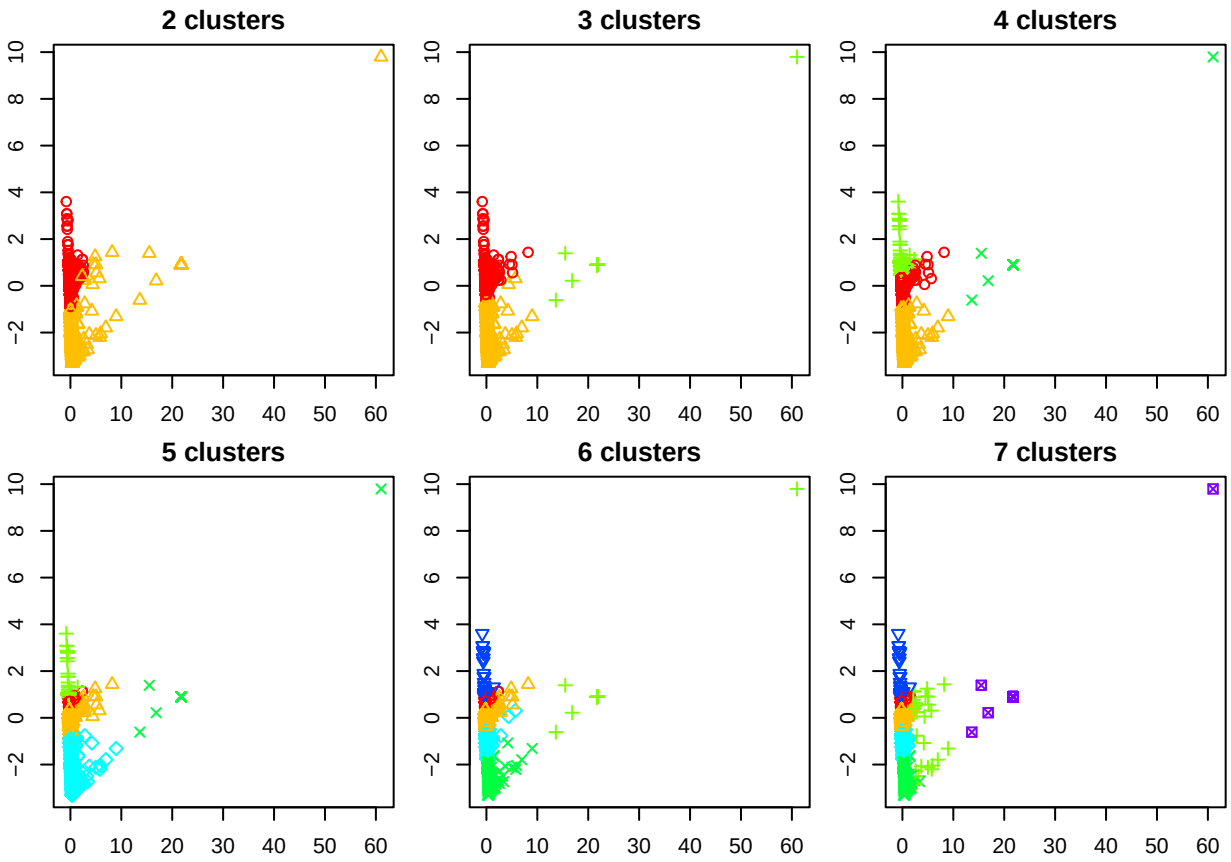





```
best_k <- 3
```

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

```
plot_clusters(brooklyn_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(brooklyn_pca$x, centers = best_k, nstart = 25)
summarize_kmeans(km, "Brooklyn")
```

```
##
## ===== K-means Model Performance Summary for Brooklyn =====
## Total within-cluster sum of squares (WSS): 4916.653
##
## Cluster sizes:
## [1] 5782   38 1343
##
## Cluster centers (in PCA space):
##   TOTAL.UNITS YEAR.BUILT GROSS.SQUARE.FEET   TOTAL.UNITS
## 1  -0.1975400  0.3775332      8.702766e-05 -0.0005465221
## 2  22.2921270  1.0924846      2.956825e-01  0.0207292682
## 3   0.2197138 -1.6563003     -8.740976e-03  0.0017664027
```

Interpret what the clusters mean with the original data

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

```
##   cluster BOROUGH RESIDENTIAL.UNITS TOTAL.UNITS GROSS.SQUARE.FEET YEAR.BUILT
## 1       1       3          2.584054    2.824628          3404.026   1918.843
```

```
## 2       2       3        469.973684  470.184211        482817.132   2014.447
## 3       3       3          3.019360    3.390916          5071.672   1977.520
```

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(brooklyn_data,
          file = "after_cluster_dataset/cleaned_brooklyn_with_clusters.csv",
          row.names = FALSE)
```