Stat4601_Manhattan_Kmeans

Stephanie Cheng

2025-04-09

K-mean with PCA

Reduce dimensions and prepare data for clustering

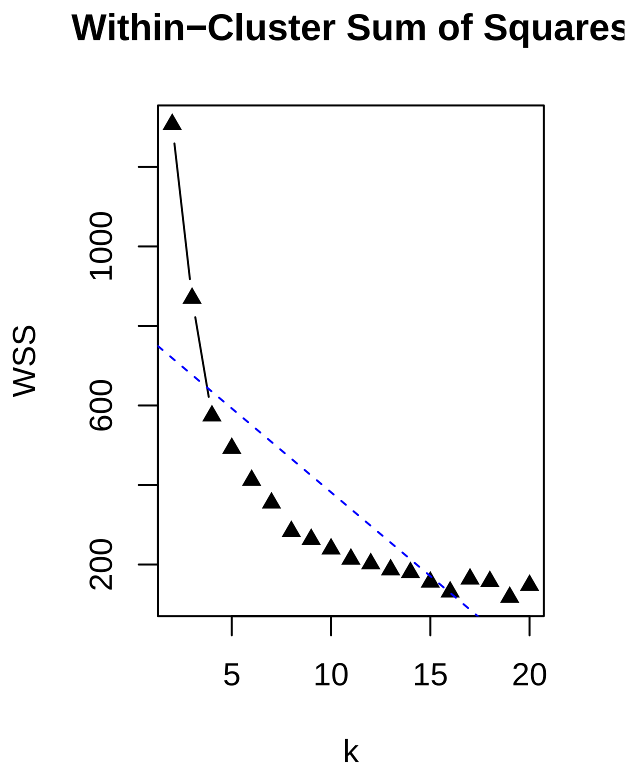
```
manhattan_pca <- pca(manhattan_data, "Manhattan")
```

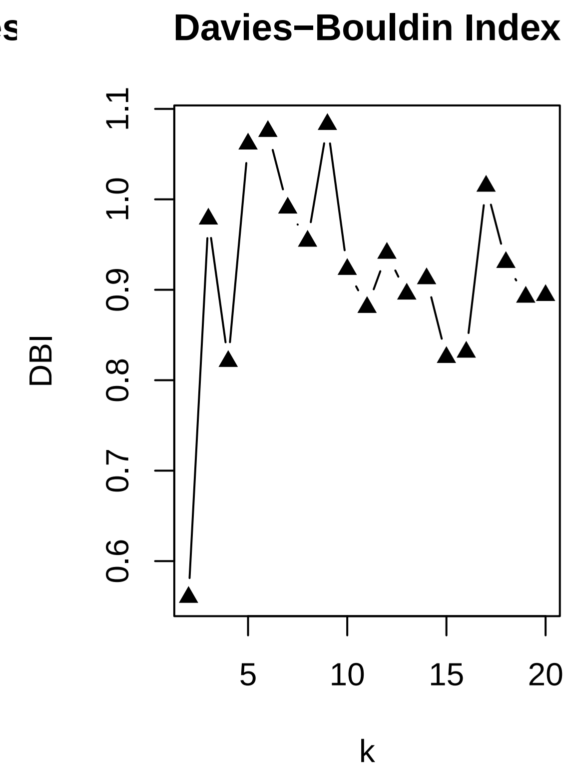
```
##
## ===== PCA Summary for Manhattan =====
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7749 0.6671 0.61683 0.15533
## Proportion of Variance 0.7876 0.1112 0.09512 0.00603
## Cumulative Proportion  0.7876 0.8989 0.99397 1.00000
## Contributing variable for each PC:
##                PC1                PC2                PC3                PC4
##       "TOTAL.UNITS" "GROSS.SQUARE.FEET"       "YEAR.BUILT"      "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.

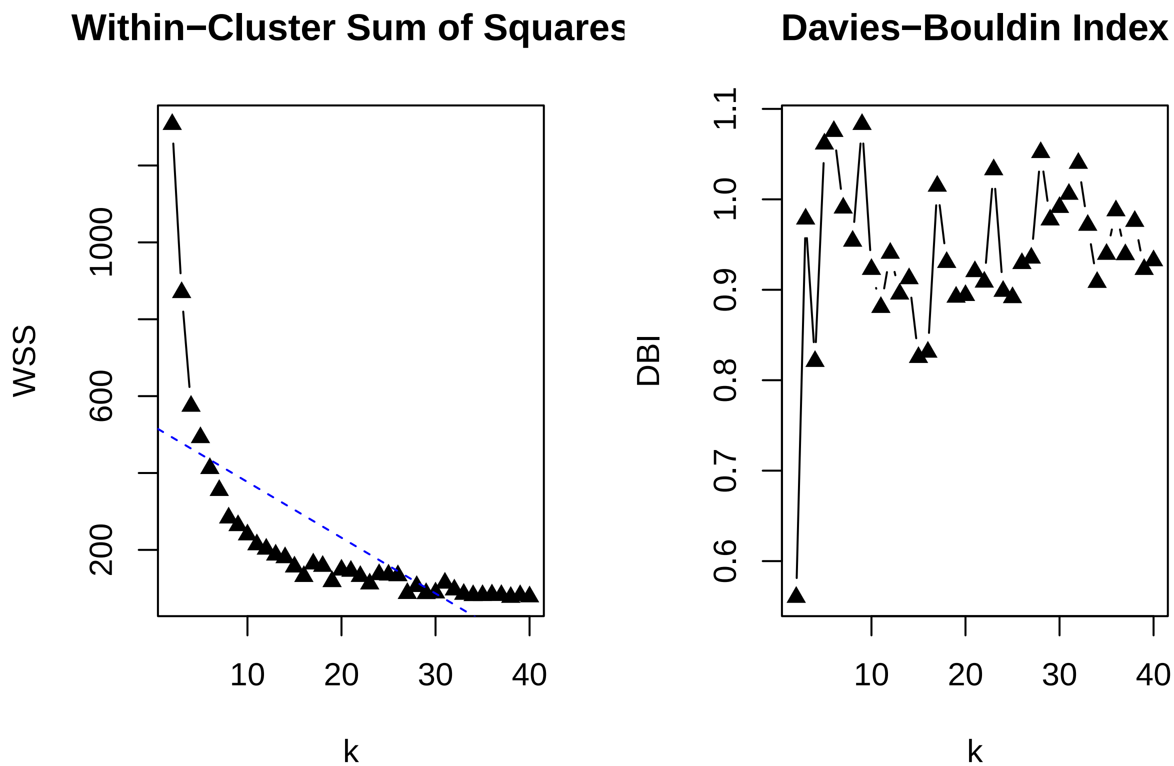
```
manhattan_k_stats_20 <- calculate_k_stats_PCA(manhattan_pca, max_k = 20)
manhattan_k_stats_40 <- calculate_k_stats_PCA(manhattan_pca, max_k = 40)

# DBI & WSS plot
elbows_20 <- plot_kmeans(manhattan_k_stats_20$errs, manhattan_k_stats_20$DBI)
```





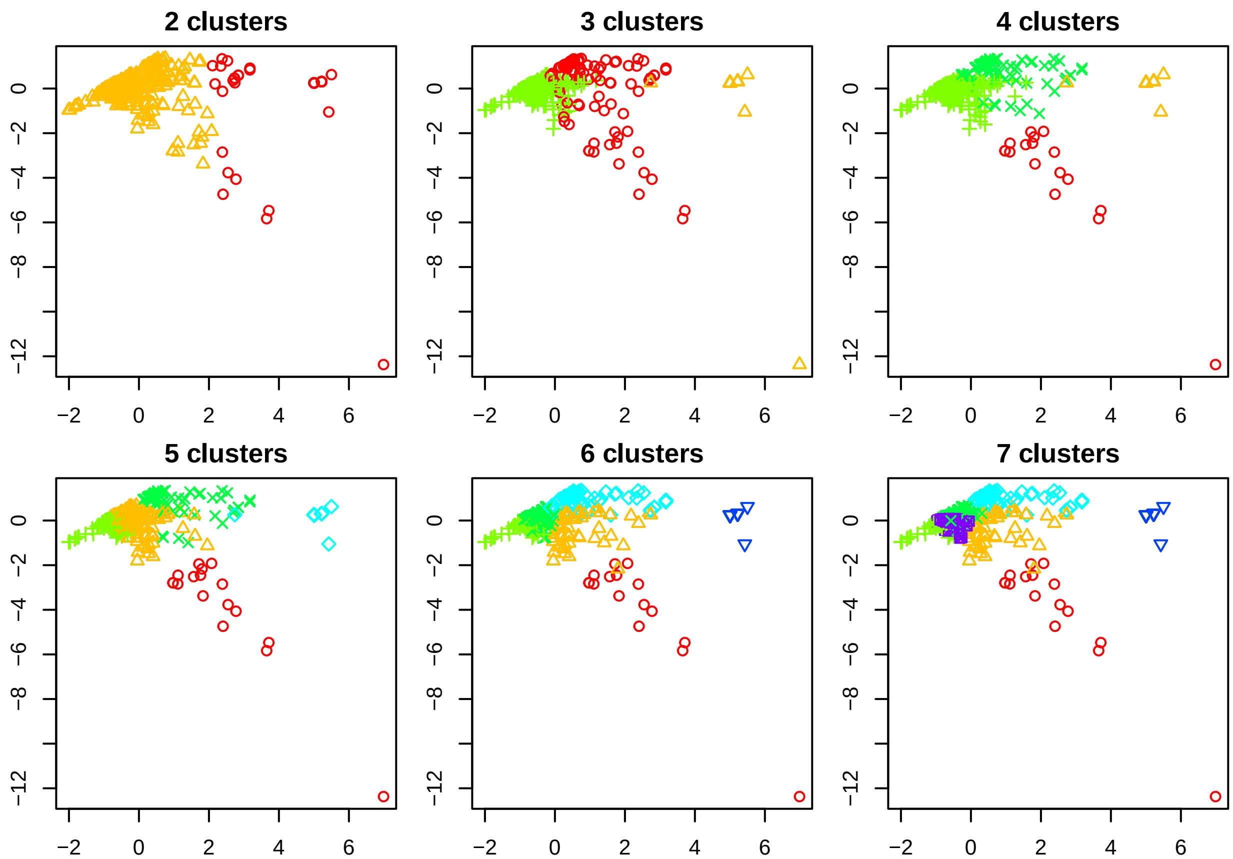

```
elbows_40 <- plot_kmeans(manhattan_k_stats_40$errs, manhattan_k_stats_40$DBI)
```




```
best_k <- 4
```

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

```
plot_clusters(manhattan_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(manhattan_pca$x, centers = best_k, nstart = 25)
summarize_kmeans(km, "Manhattan")
```

```
##
## ===== K-means Model Performance Summary for Manhattan =====
## Total within-cluster sum of squares (WSS): 573.9247
##
## Cluster sizes:
## [1]  18  89  99 870
##
## Cluster centers (in PCA space):
##   TOTAL.UNITS GROSS.SQUARE.FEET YEAR.BUILT  TOTAL.UNITS
## 1   2.2425040       -3.72981615 -1.4336569  0.047140474
## 2   0.9631460        0.70608247 -1.4291193  0.054834647
## 3   5.0649792        0.26087147  0.5281862 -0.047778545
## 4  -0.7212851       -0.02474819  0.1157552 -0.001147973
```

Interpret what the clusters mean with the original data

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

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

```
## 1       1       1          0.000000    63.05556          722848.00   1948.000
## 2       2       1         37.146067    47.88764           80640.88   1993.652
## 3       3       1        281.727273   283.33333          355294.00   2001.384
## 4       4       1          9.524138    12.40805           18450.32   1910.329
```

Export the clusters for Supervised learning

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

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

```r
write.csv(manhattan_data,
          file = "after_cluster_dataset/cleaned_manhattan_with_clusters.csv",
          row.names = FALSE)
```