

# Statistical Learning Workshop - K-Means

Abdollah Safari

26/02/2021

## Load data

For all demonstrations of clustering, we'll use the `penguins` dataset, which is available in the R package `palmerpenguins` and is also described at <https://bit.ly/2N9pIB9> (<https://bit.ly/2N9pIB9>)

```
p_data(palmerpenguins)           # Info on two datasets
```

```
##      Data                Description
## 1 penguins              Size measurements for adult foraging penguins near Palmer
##                               r Station, Antarctica
## 2 penguins_raw (penguins) Penguin size, clutch, and blood isotope data for foraging
##                               adults near Palmer Station, Antarctica
```

```
?palmerpenguins::penguins        # Get help with links
palmerpenguins::penguins         # See first dataset
```

```
## # A tibble: 344 x 8
##   species island bill_length_mm bill_depth_mm flipper_length... body_mass_g
##   <fct>   <fct>         <dbl>         <dbl>         <int>         <int>
## 1 Adelie Torge...      39.1          18.7          181          3750
## 2 Adelie Torge...      39.5          17.4          186          3800
## 3 Adelie Torge...      40.3           18           195          3250
## 4 Adelie Torge...      NA            NA            NA            NA
## 5 Adelie Torge...      36.7          19.3          193          3450
## 6 Adelie Torge...      39.3          20.6          190          3650
## 7 Adelie Torge...      38.9          17.8          181          3625
## 8 Adelie Torge...      39.2          19.6          195          4675
## 9 Adelie Torge...      34.1          18.1          193          3475
## 10 Adelie Torge...      42            20.2          190          4250
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

```
palmerpenguins::penguins_raw     # See second dataset
```

```
## # A tibble: 344 x 17
##   studyName `Sample Number` Species Region Island Stage `Individual ID`
##   <chr>          <dbl> <chr>   <chr>   <chr>   <chr> <chr>
## 1 PAL0708          1 Adelie... Anvers Torge... Adul... N1A1
## 2 PAL0708          2 Adelie... Anvers Torge... Adul... N1A2
## 3 PAL0708          3 Adelie... Anvers Torge... Adul... N2A1
## 4 PAL0708          4 Adelie... Anvers Torge... Adul... N2A2
## 5 PAL0708          5 Adelie... Anvers Torge... Adul... N3A1
## 6 PAL0708          6 Adelie... Anvers Torge... Adul... N3A2
## 7 PAL0708          7 Adelie... Anvers Torge... Adul... N4A1
## 8 PAL0708          8 Adelie... Anvers Torge... Adul... N4A2
## 9 PAL0708          9 Adelie... Anvers Torge... Adul... N5A1
## 10 PAL0708         10 Adelie... Anvers Torge... Adul... N5A2
## # ... with 334 more rows, and 10 more variables: `Clutch Completion` <chr>, `Date
## #   Egg` <date>, `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>, `Delta 15 N
## #   (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

```
(df <- palmerpenguins::penguins) # Save/show data to df
```

```
## # A tibble: 344 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_... body_mass_g
##   <fct>   <fct>          <dbl>          <dbl>          <int>          <int>
## 1 Adelie Torge...        39.1           18.7           181           3750
## 2 Adelie Torge...        39.5           17.4           186           3800
## 3 Adelie Torge...        40.3           18            195           3250
## 4 Adelie Torge...        NA            NA            NA            NA
## 5 Adelie Torge...        36.7           19.3           193           3450
## 6 Adelie Torge...        39.3           20.6           190           3650
## 7 Adelie Torge...        38.9           17.8           181           3625
## 8 Adelie Torge...        39.2           19.6           195           4675
## 9 Adelie Torge...        34.1           18.1           193           3475
## 10 Adelie Torge...        42            20.2           190           4250
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

## Prepare data

```
df %>% summary()
```

```
##           species           island  bill_length_mm  bill_depth_mm
## Adelie      :152  Biscoe      :168  Min.      :32.10  Min.      :13.10
## Chinstrap: 68  Dream       :124  1st Qu.:39.23  1st Qu.:15.60
## Gentoo     :124  Torgersen: 52  Median :44.45  Median :17.30
##                                     Mean   :43.92  Mean    :17.15
##                                     3rd Qu.:48.50  3rd Qu.:18.70
##                                     Max.   :59.60  Max.    :21.50
##                                     NA's   :2      NA's    :2
## flipper_length_mm  body_mass_g      sex      year
## Min.      :172.0    Min.      :2700  female:165  Min.      :2007
## 1st Qu.:190.0    1st Qu.:3550  male  :168  1st Qu.:2007
## Median :197.0    Median :4050  NA's   : 11  Median :2008
## Mean     :200.9    Mean     :4202                Mean     :2008
## 3rd Qu.:213.0    3rd Qu.:4750                3rd Qu.:2009
## Max.     :231.0    Max.     :6300                Max.     :2009
## NA's     :2      NA's     :2
```

```
# Select and rename variables
df %<>%
  as_tibble()%>%
  rename(y = species) %>% # Rename species to y
  select(                  # Use `select` to remove vars
    -island,               # Remove island
    -sex,                  # Remove sex
    -year) %>%            # Remove year
  na.omit()                # Remove incomplete cases

# Look at the first few rows of the prepared data frame
df
```

```
## # A tibble: 342 x 5
##   y      bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>          <dbl>         <dbl>          <int>         <int>
## 1 Adelie          39.1           18.7            181           3750
## 2 Adelie          39.5           17.4            186           3800
## 3 Adelie          40.3           18              195           3250
## 4 Adelie          36.7           19.3            193           3450
## 5 Adelie          39.3           20.6            190           3650
## 6 Adelie          38.9           17.8            181           3625
## 7 Adelie          39.2           19.6            195           4675
## 8 Adelie          34.1           18.1            193           3475
## 9 Adelie          42              20.2            190           4250
## 10 Adelie         37.8           17.1            186           3300
## # ... with 332 more rows
```

```
# SAVE DATA #####

# Use saveRDS(), which saves data to native R formats
df %>% saveRDS("~/Downloads/Clustering/Data/penguins.rds")

# Set random seed for reproducibility in processes like
# splitting the data
set.seed(1) # You can use any number here

# Import data and sample
df <- import("~/Downloads/Clustering/Data/penguins.rds") %>%
  sample_n(100) # Reduce n for graphing

# Look at the first few rows of the prepared data frame
df
```

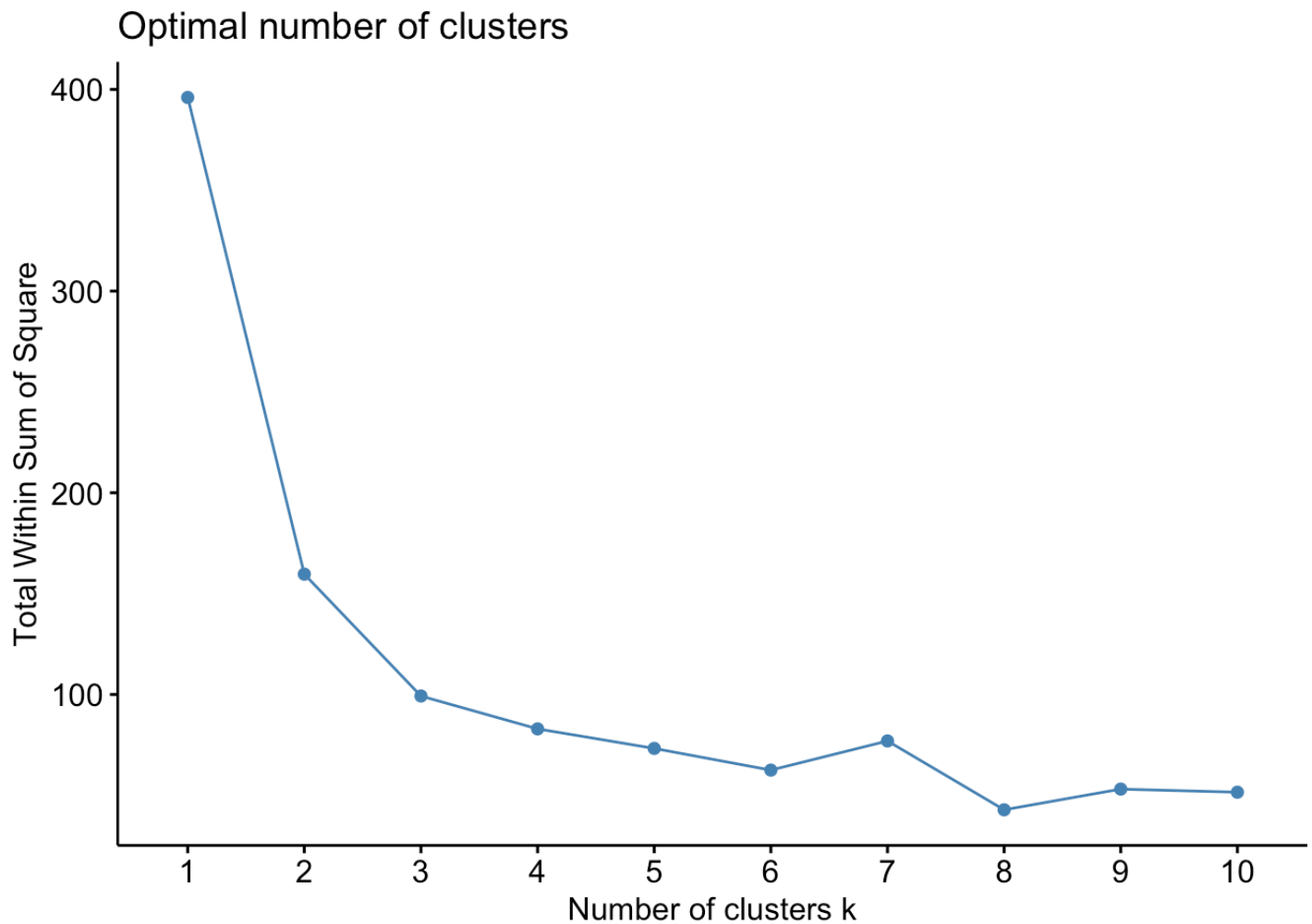
```
## # A tibble: 100 x 5
##   y          bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>          <dbl>          <dbl>          <int>          <int>
## 1 Chinstrap      49.8            17.3            198            3675
## 2 Gentoo         49.3            15.7            217            5850
## 3 Adelie         44.1            18              210            4000
## 4 Chinstrap      46.7            17.9            195            3300
## 5 Gentoo         47.2            13.7            214            4925
## 6 Gentoo         48.4            16.3            220            5400
## 7 Chinstrap      42.5            16.7            187            3350
## 8 Adelie         41.3            20.3            194            3550
## 9 Chinstrap      51.3            19.2            193            3650
## 10 Chinstrap     52.2            18.8            197            3450
## # ... with 90 more rows
```

```
# Separate the class labels
species <- df %>% # Rename `y` back to `species`
  pull(y)        # Select just `y` as a vector

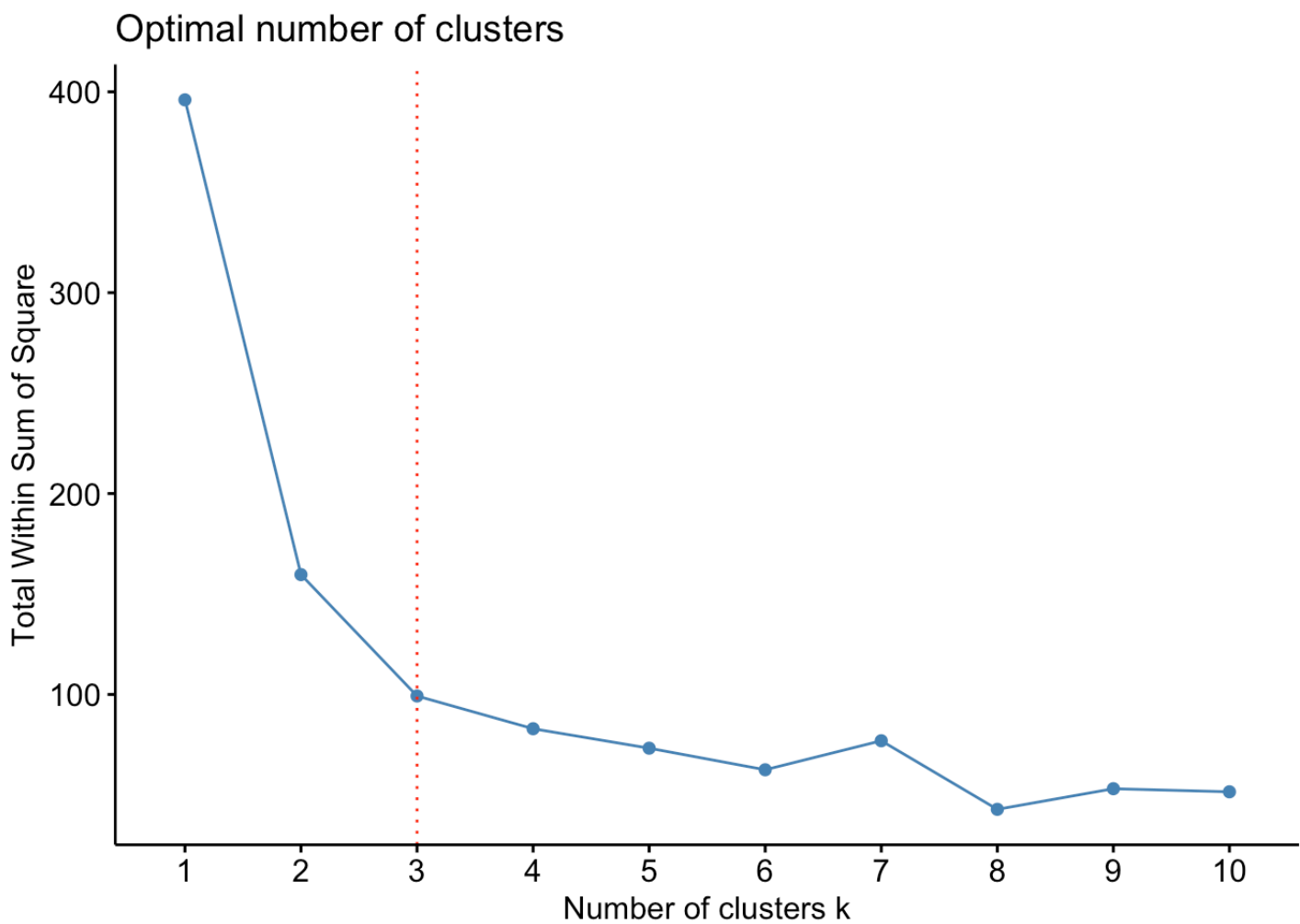
df %<>%
  select(-y) %>% # Select everything except `y`
  scale()        # Standardize variables
```

## Number of clusters (K)

```
# OPTIMAL NUMBER OF CLUSTERS #####  
  
# Elbow method  
df %>%  
  fviz_nbclust(      # From `factoextra`  
    FUN = kmeans,    # Use k-means  
    method = "wss"    # "within cluster sums of squares"  
  )
```

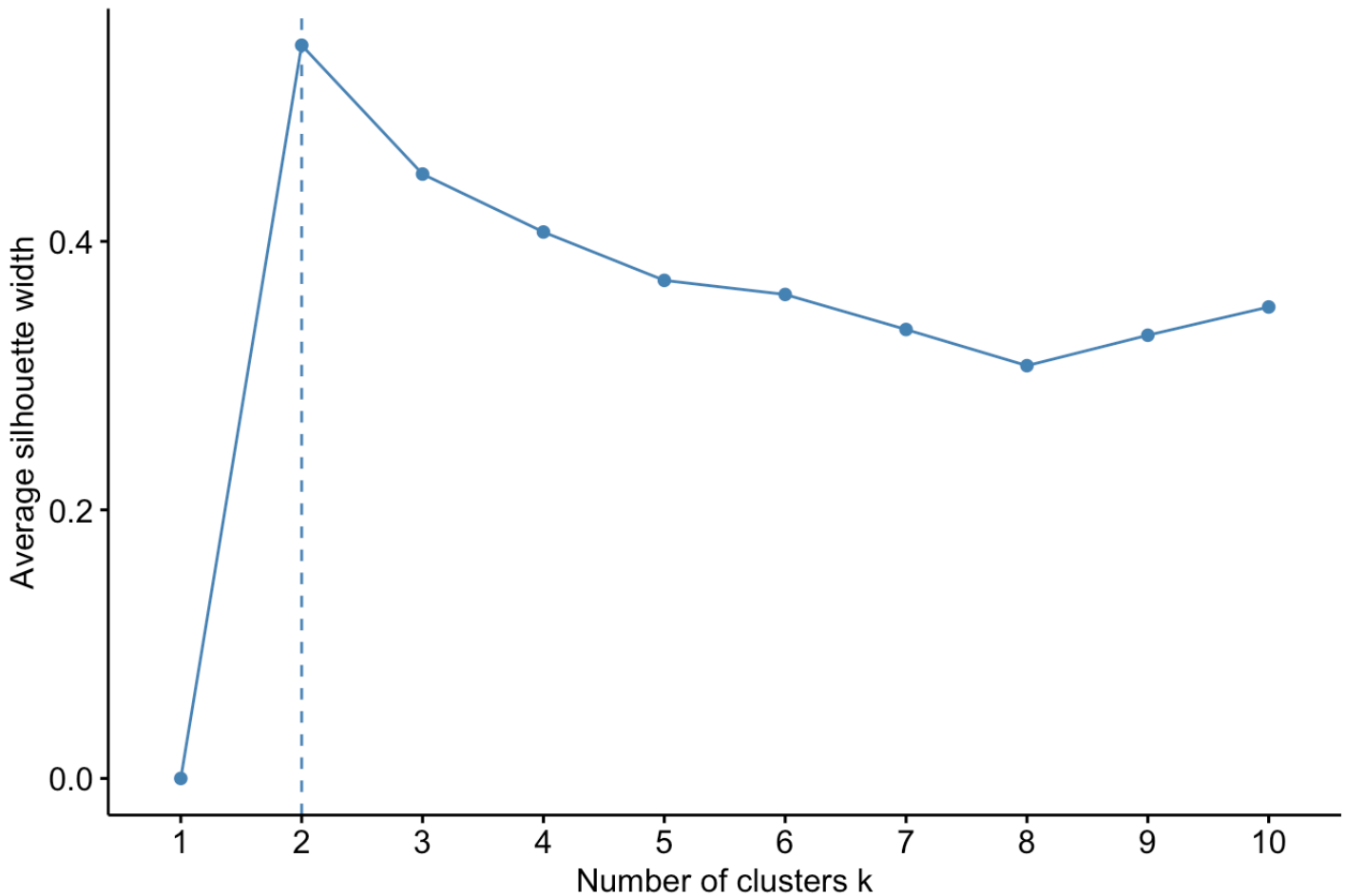


```
# Adding a reference line
df %>%
  fviz_nbclust(          # From `factoextra`
    FUN = kmeans,        # Use k-means
    method = "wss"       # "within cluster sums of squares"
  ) +
  geom_vline(            # Reference line
    xintercept = 3,
    color = "red",
    linetype = "dotted"
  )                       # Look for "bend" in curve
```

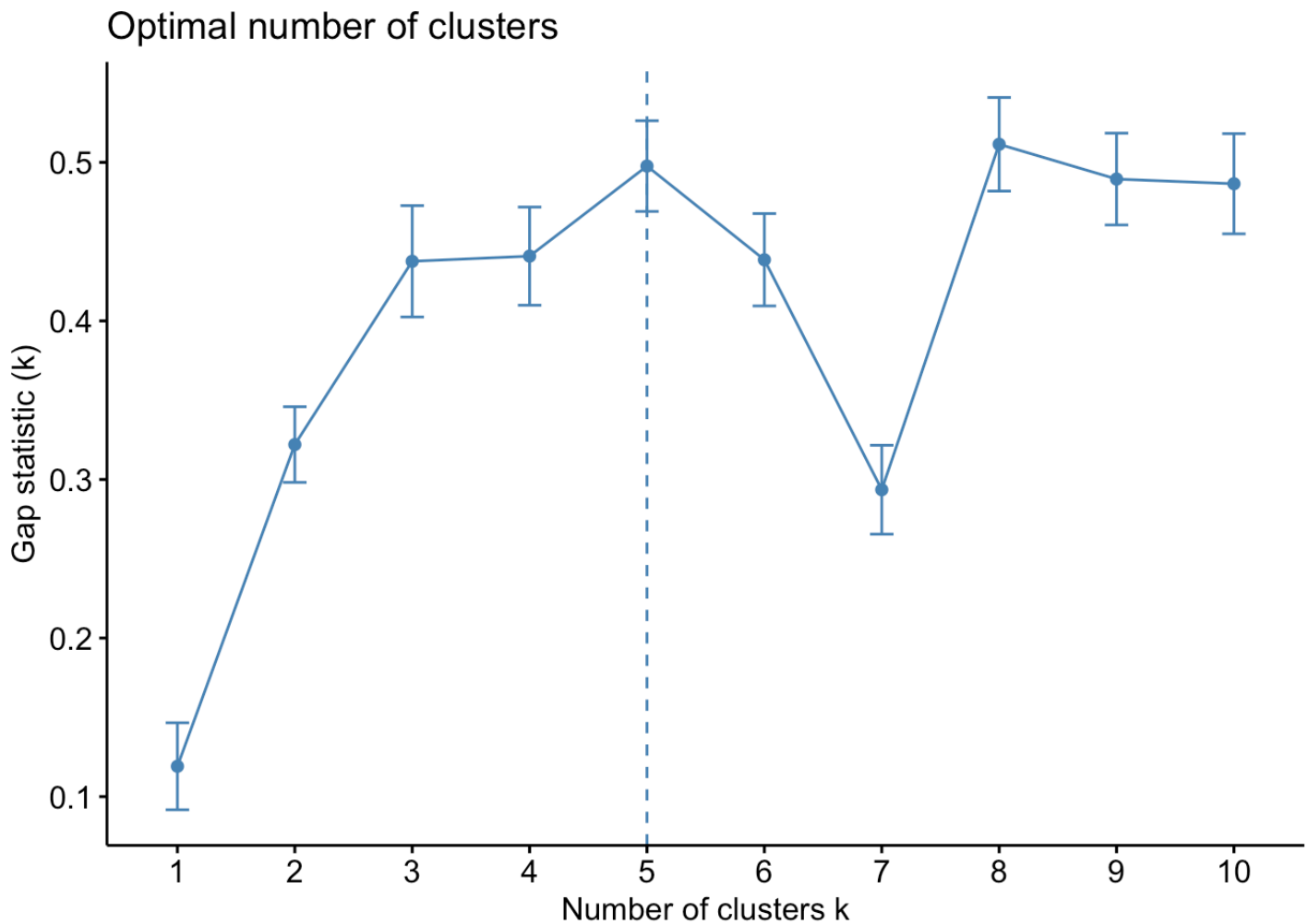


```
# Silhouette method
df %>%
  fviz_nbclust(
    FUN = kmeans,        # Use k-means
    method = "silhouette" # Look for maximum width
  )
```

## Optimal number of clusters



```
# Use gap statistics to find optimal number of clusters
# and visualize it using fviz_gap_stat
df %>%
  clusGap(          # Function from `cluster`
    FUN = kmeans,   # Method for clustering
    K.max = 10,     # Maximum number of clusters
    B = 100         # Number of Monte Carlo/bootstrap samples
  ) %>%
  fviz_gap_stat()   # Look for highest value
```



## K-means clustering

```
# Compute three clusters
km <- df %>%
  kmeans(3) %>% # Set the number of clusters
  print()       # Print output
```



```
## K-means clustering with 3 clusters of sizes 70, 13, 17
##
## Cluster means:
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1      -0.3375706      0.5298954      -0.5716827    -0.5398149
## 2       0.4411131     -1.6277604       1.0292951     0.6615704
## 3       1.0526748     -0.9371645       1.5668795     1.7168603
##
## Clustering vector:
##   [1] 1 3 1 1 2 3 1 1 1 1 3 1 1 3 1 1 1 1 1 1 1 1 1 3 1 3 2 1 1 1 2 1
##  [38] 1 1 1 2 2 3 1 3 1 1 1 1 2 1 1 1 1 2 3 1 1 1 1 1 1 1 3 1 1 1 1 1 2
##  [75] 1 2 1 3 2 2 3 1 1 1 1 1 3 1 1 2 1 2 1 3 1 1 1 3 1 1
##
## Within cluster sum of squares by cluster:
## [1] 131.372672   4.726664   6.941080
## (between_SS / total_SS =  63.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# Visualize the clusters
km %>% fviz_cluster(
  data = df,
  geom = c("point")
) +
  geom_text(
    vjust = 1.5, # Color points according to cluster
    aes(
      color = as.factor(km$cluster),
      label = species # label according to species
    )
  )
)
```

## Cluster plot

