lab01

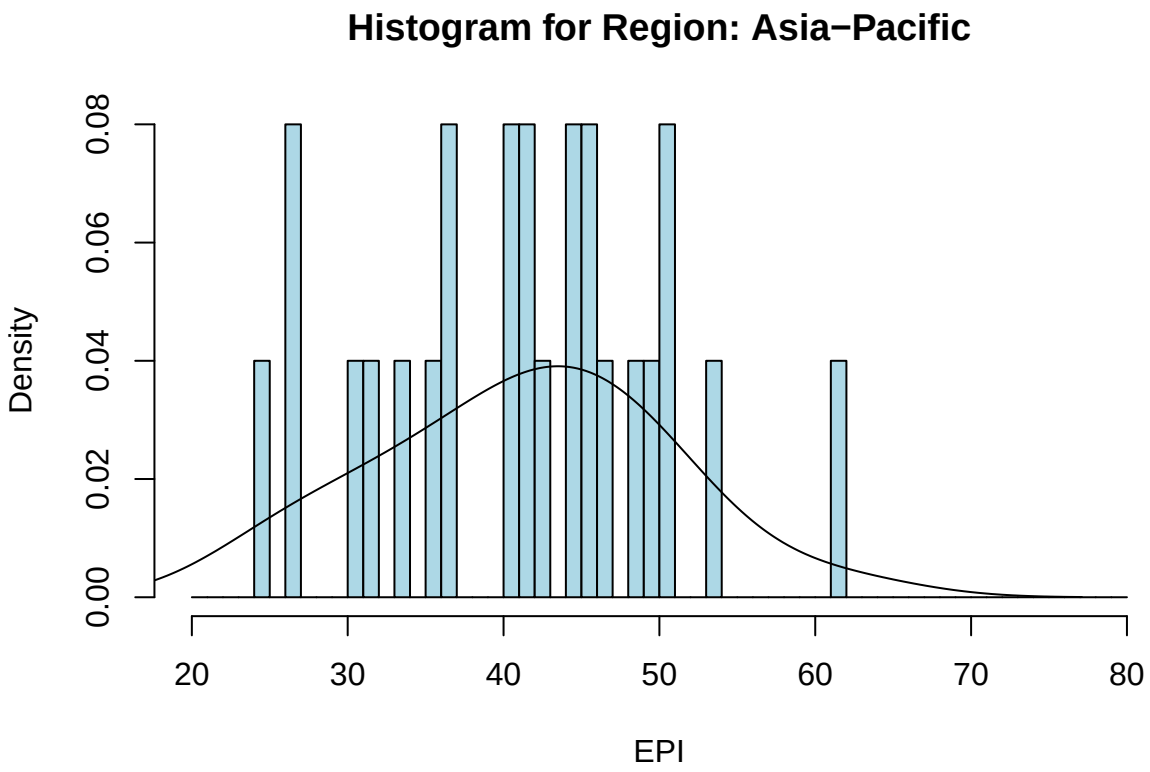
Sean Fitch

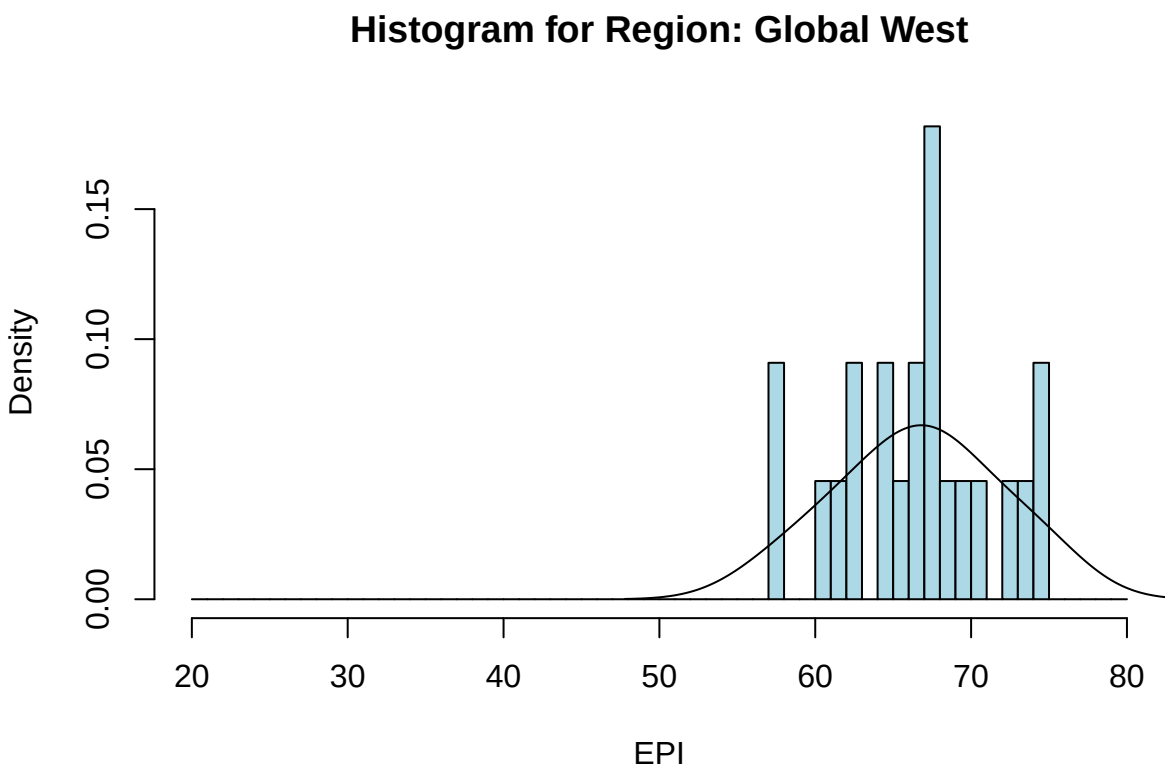
2024-10-04

Load the results dataset

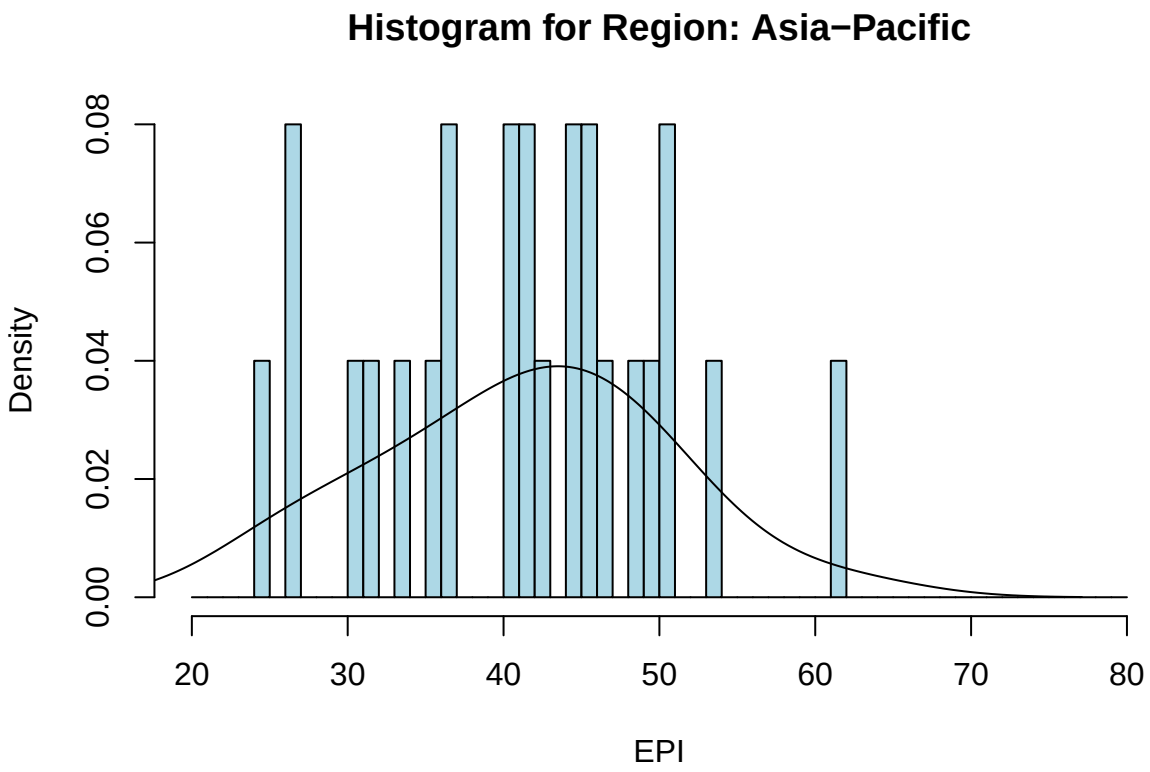
```
# read data
epi <- read.csv("../data/epi2024results_DA_F24_lab03.csv")
```

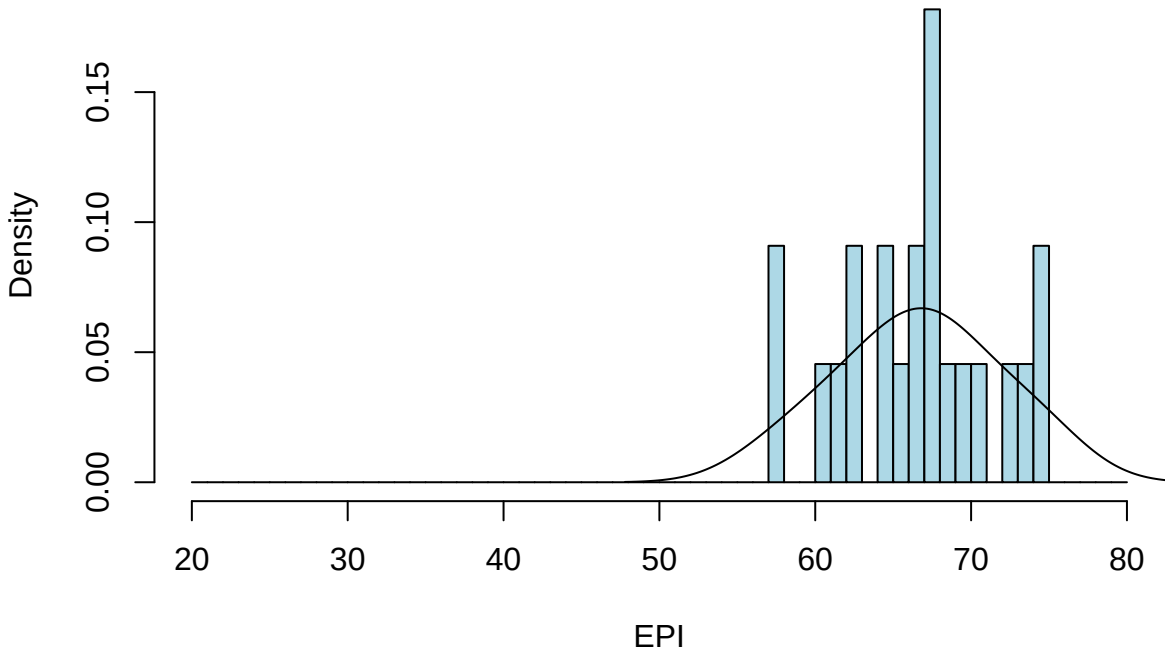
```
selected_regions <- epi %>%
  distinct(region) %>%
  slice_tail(n = 2) %>%
  pull(region)

epi %>%
  filter(region %in% selected_regions) %>%
  group_by(region) %>%
  group_walk(~ {
    region_name <- .y$region
        EPI_data <- .x$EPI

    # Create a histogram
    hist(EPI_data, breaks = seq(20, 80, 1.0), prob = TRUE,
         main = paste("Histogram for Region:", region_name),
         xlab = "EPI", ylab = "Density", col = "lightblue", border = "black")

    # Add the density line
    lines(density(EPI_data, na.rm = TRUE, bw = 'SJ'))
  })
```


Histogram for Region: Asia−Pacific
Density
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
70
80
EPI




```
epi %>%
  filter(region %in% selected_regions) %>%
  group_by(region) %>%
  group_walk(~ {
    region_name <- .y$region
    EPI_data <- .x$EPI

    # Create a histogram
    hist(EPI_data, breaks = seq(20, 80, 1.0), prob = TRUE,
         main = paste("Histogram for Region:", region_name),
         xlab = "EPI", ylab = "Density", col = "lightblue", border = "black")

    # Add the density line
    lines(density(EPI_data, na.rm = TRUE, bw = 'SJ'))
  })
```


Histogram for Region: Asia−Pacific
Density
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
70
80
EPI

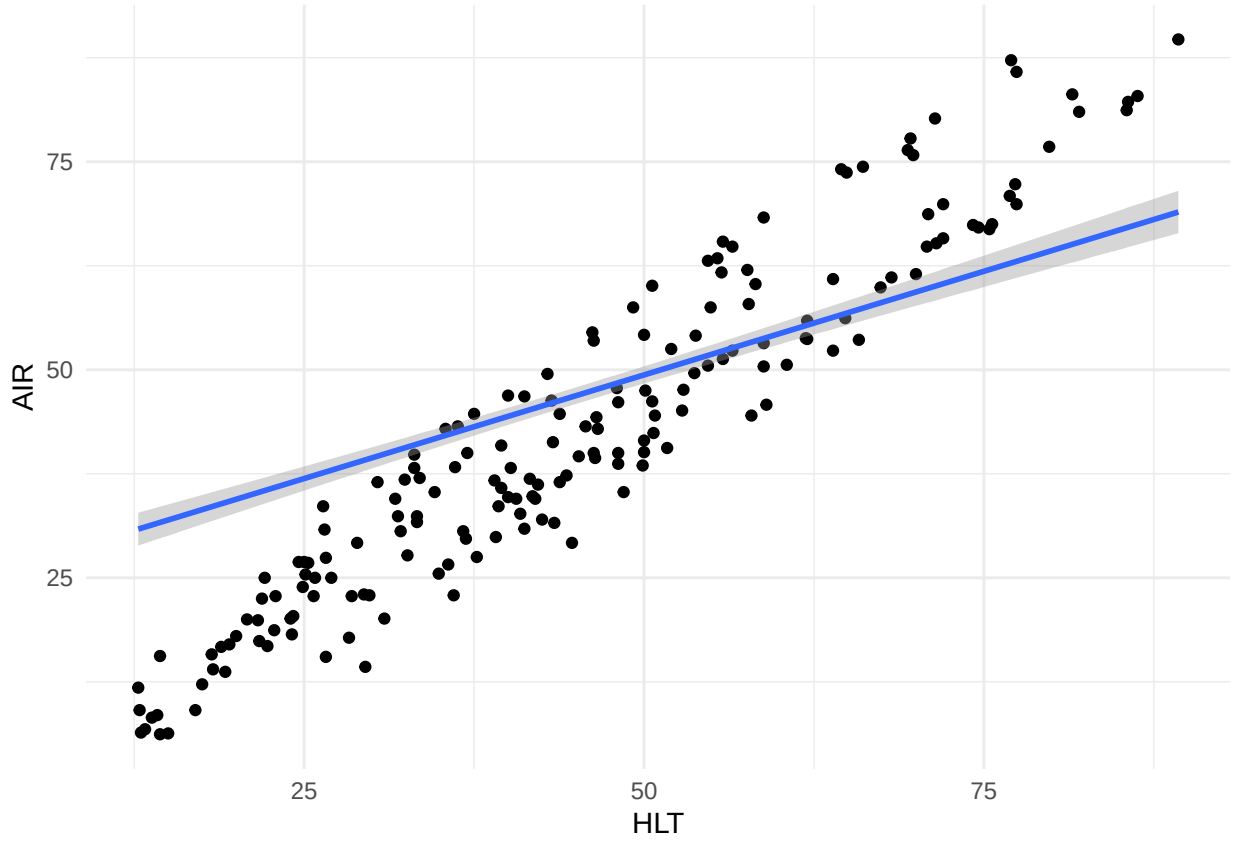



```r
linear_model <- lm(EPI ~ WWT + WWR + HLT + AIR + HPE, data = epi)
summary(linear_model)
```

```
##
## Call:
## lm(formula = EPI ~ WWT + WWR + HLT + AIR + HPE, data = epi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5124  -4.4814   0.6935   4.2839  18.2293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.64374    1.41326  17.438  < 2e-16 ***
## WWT          0.04772    0.02855   1.672   0.0964 .
## WWR          0.02863    0.03665   0.781   0.4358
## HLT          0.75468    0.14230   5.304 3.42e-07 ***
## AIR         -0.36305    0.16481  -2.203   0.0289 *
## HPE          0.01975    0.04835   0.409   0.6834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.048 on 174 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.7259
## F-statistic: 95.79 on 5 and 174 DF,  p-value: < 2.2e-16
```

```r
ggplot(epi, aes(x = HLT, y = AIR)) +
  geom_point() +
  stat_smooth(method = "lm", aes(y = EPI)) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



HLT most significantly influences EPI.

```r
sub_saharan_africa_epi <- epi %>%
  filter(region == "Sub-Saharan Africa")
linear_model <- lm(EPI ~ WWT + WWR + HLT + AIR + HPE, data = sub_saharan_africa_epi)
summary(linear_model)
```

```
##
## Call:
## lm(formula = EPI ~ WWT + WWR + HLT + AIR + HPE, data = sub_saharan_africa_epi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4780 -2.7365 -0.2571  2.1588 10.6573
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.16511    2.56907  14.856   <2e-16 ***
```

```
## WWT          0.07774    0.04585   1.695   0.0978 .
## WWR          0.20308    0.09536   2.130   0.0394 *
## HLT         -0.14346    0.46958  -0.306   0.7616
## AIR          0.07154    0.49606   0.144   0.8861
## HPE          0.01275    0.09763   0.131   0.8968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.273 on 40 degrees of freedom
## Multiple R-squared:  0.4241, Adjusted R-squared:  0.3521
## F-statistic: 5.891 on 5 and 40 DF,  p-value: 0.0003606
```

The full model is a better fit. It captures more variance, as shown by and Adjusted R-Squared of 0.7259 versus 0.3521. It has higher residuals, but that can be explained by the variance of EPI being higher. It is also worth noting that the significance of predictors varies, implying there are different relationships between the predictors and EPI in different regions.

```
epi.norm <- epi %>%
  select(EPI, WWT, WWR, HLT, AIR)
epi.norm <- as.data.frame(scale(epi.norm))
epi.norm$region <- epi$region

# Define the regions
regions_a <- c("Latin America & Caribbean", "Asia-Pacific", "Eastern Europe")
regions_b <- c("Sub-Saharan Africa", "Global West", "Greater Middle East")
epi.norm.a <- epi.norm %>%
  filter(region %in% regions_a)
epi.norm.b <- epi.norm %>%
  filter(region %in% regions_b)

# Create a random sample of 80% of the data
set.seed(123)  # Set seed for reproducibility
sample_index.a <- sample(seq_len(nrow(epi.norm.a)), size = 0.8 * nrow(epi.norm.a))
sample_index.b <- sample(seq_len(nrow(epi.norm.b)), size = 0.8 * nrow(epi.norm.b))
train_a <- epi.norm.a[sample_index.a, ]
test_a <- epi.norm.a[-sample_index.a, ]
train_b <- epi.norm.b[sample_index.b, ]
test_b <- epi.norm.b[-sample_index.b, ]

KNNpred.a <- knn(train = train_a[1:5], test = test_a[1:5], cl = train_a$region)
KNNpred.b <- knn(train = train_b[1:5], test = test_b[1:5], cl = train_b$region)
```

```
contingency_a <- table(Predicted = KNNpred.a, Actual = test_a$region)
contingency_b <- table(Predicted = KNNpred.b, Actual = test_b$region)

contingency_a
```

```
##                            Actual
## Predicted                   Asia-Pacific Eastern Europe
##   Asia-Pacific                         2              1
##   Eastern Europe                       1              1
##   Latin America & Caribbean            2              2
##                            Actual
```

```
## Predicted                   Latin America & Caribbean
##   Asia-Pacific                                     1
##   Eastern Europe                                   0
##   Latin America & Caribbean                        6
```

```
contingency_b
```

```
##                      Actual
## Predicted             Global West Greater Middle East Sub-Saharan Africa
##   Global West                   2                   0                  0
##   Greater Middle East           0                   5                  0
##   Sub-Saharan Africa            0                   1                  9
```

```
accuracy_a <- sum(diag(contingency_a)) / sum(contingency_a)
accuracy_b <- sum(diag(contingency_b)) / sum(contingency_b)
print(paste("Accuracy a:", round(accuracy_a * 100, 2)))
```

```
## [1] "Accuracy a: 56.25"
```

```
print(paste("Accuracy b:", round(accuracy_b * 100, 2)))
```

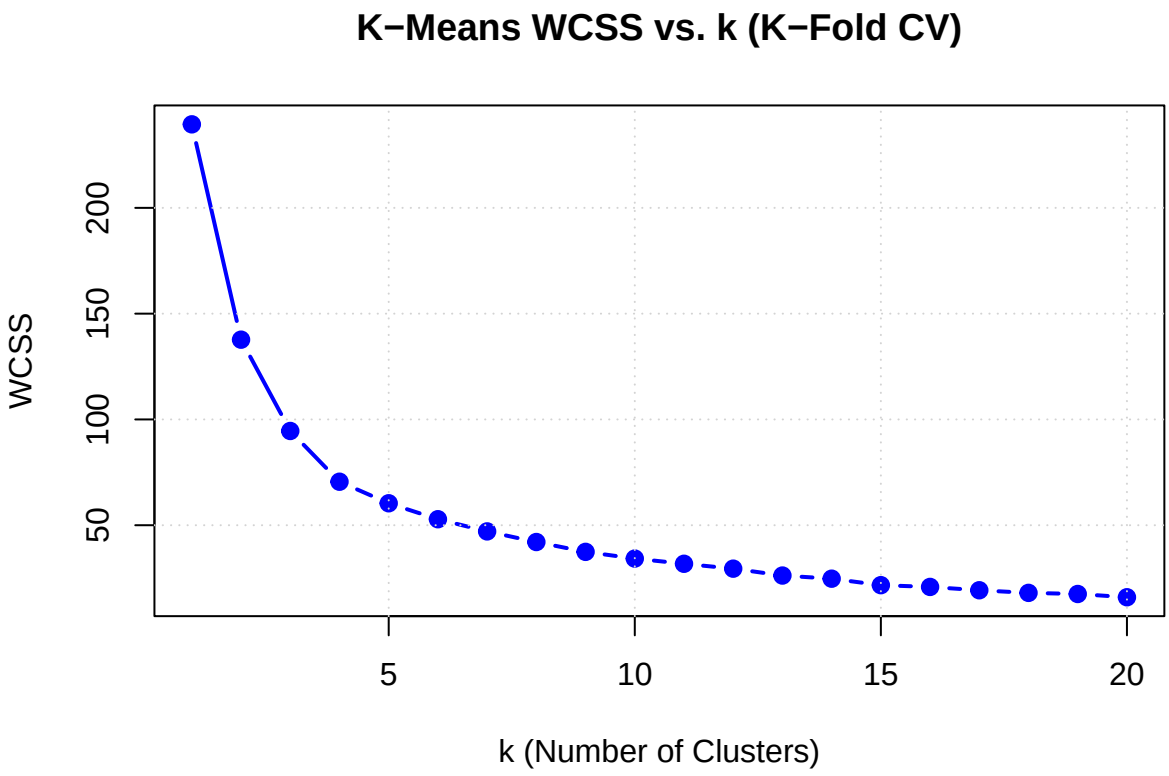
```
## [1] "Accuracy b: 94.12"
```

The accuracy for region b is significantly greater. This is likely due to greater differences in the chosen columns between the regions chosen in the group.

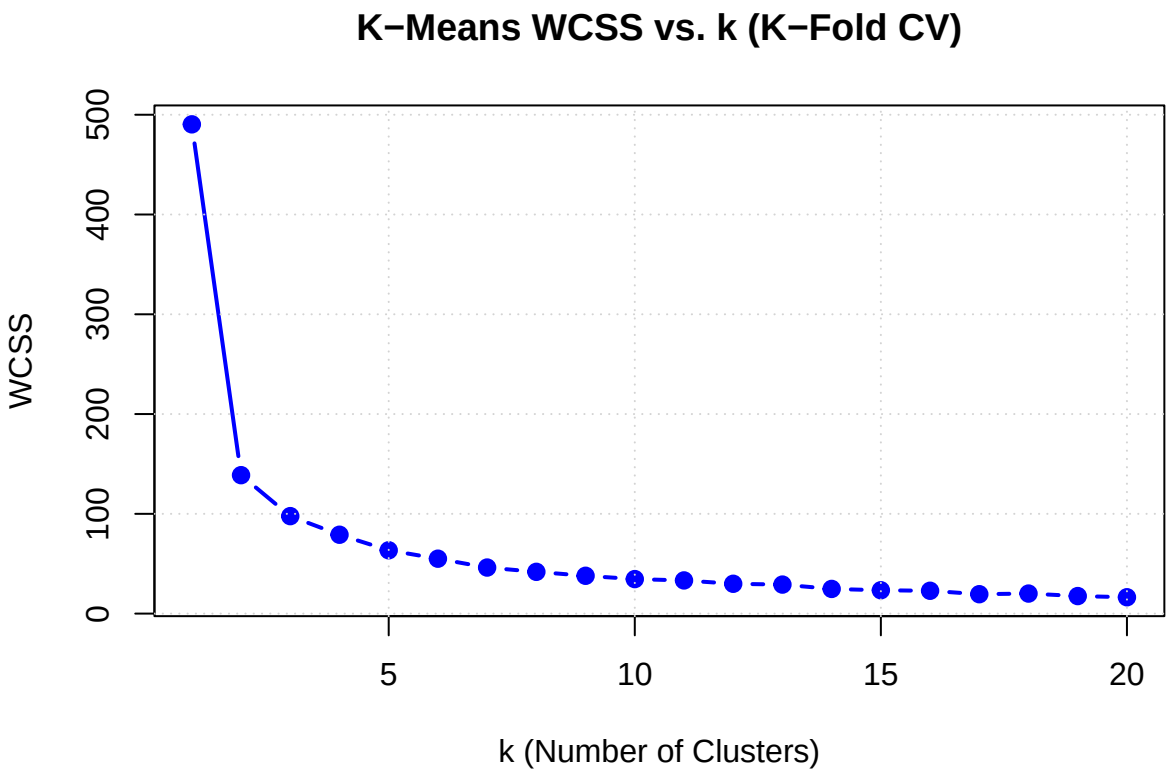
```
k_fold_kmeans <- function(data, k_values = seq(1, 20), k_folds = 10, seed = 123) {
  # Set the seed for reproducibility
  set.seed(seed)

  # Create a k-fold cross-validation partition
  folds <- createFolds(data$region, k = k_folds, list = FALSE)

  # Initialize a vector to store WCSS for different k values
  wcss_values <- numeric(length(k_values))

  # Loop through different values of k
  for (i in seq_along(k_values)) {
    k <- k_values[i]  # Get the current value of k
    fold_wcss <- numeric(k_folds)  # Initialize vector to store WCSS for each fold

    # Loop through each fold
    for (fold in seq_len(k_folds)) {
      # Split the data into training and testing sets based on the fold
      train_data <- data[folds != fold, ]

      # Perform K-means clustering on the training data
      kmeans_model <- kmeans(train_data[, -ncol(train_data)], centers = k)

      # Calculate WCSS for this fold and store it
      fold_wcss[fold] <- sum(kmeans_model$withinss)
```

```
    }

    # Average WCSS across all folds for this k value
    wcss_values[i] <- mean(fold_wcss)
  }

  # Plot the WCSS vs. k values
  plot(k_values, wcss_values, type = "b", pch = 19, col = "blue", lwd = 2,
       xlab = "k (Number of Clusters)", ylab = "WCSS",
       main = "K-Means WCSS vs. k (K-Fold CV)")
  grid()  # Add grid lines
}
```

```
epi.norm <- epi %>%
  select(EPI, WWT, WWR, HLT, AIR)
epi.norm <- as.data.frame(scale(epi.norm))
epi.norm$region <- epi$region

# Define the regions
regions_a <- c("Latin America & Caribbean", "Asia-Pacific", "Eastern Europe")
regions_b <- c("Sub-Saharan Africa", "Global West", "Greater Middle East")
epi.norm.a <- epi.norm %>%
  filter(region %in% regions_a)
epi.norm.b <- epi.norm %>%
  filter(region %in% regions_b)

k_fold_kmeans(epi.norm.a)
```




```
k_fold_kmeans(epi.norm.b)
```



With group a we see tighter clusters, with a wcss of ~65 at the elbow vs. ~100 for b. This either indicates less variance in group a or better separation of clusters.