Cluster Analysis

Transorm tot.form into binary where if an individual used more than one form, the value is one, else, zero.

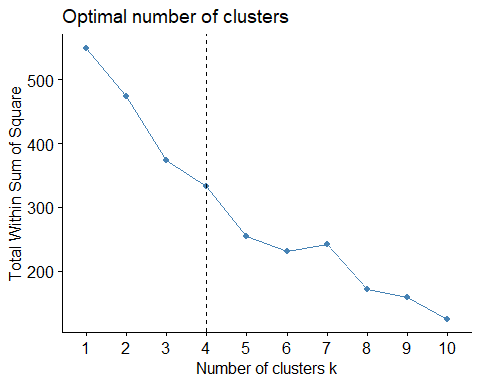
## 1. Person Accusitive Plural

# Read in the data  
data <- PerAccPlu.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -22.2961 25212.3199 -0.001 0.999  
## Gender.agefemale under 75 -19.9370 14948.4247 -0.001 0.999  
## Gender.agemale -19.5731 14585.5467 -0.001 0.999  
## AreaPoland 0.4376 20555.0831 0.000 1.000  
## AreaWest 19.6665 12648.0075 0.002 0.999  
## Length2 -18.3253 27045.7236 -0.001 0.999  
## Length3 1.2433 24018.3721 0.000 1.000  
## Length4 -19.5641 33323.3880 -0.001 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10.238 on 61 degrees of freedom  
## Residual deviance: 5.004 on 54 degrees of freedom  
## AIC: 21.004  
##   
## Number of Fisher Scoring iterations: 22

From this output, it seems that none of the predictors (Gender.age, Area, Length) are statistically significant in predicting the outcome tot.form

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



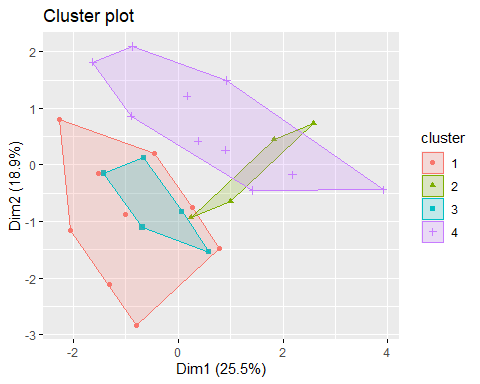
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0 16 male Poland NA  
## 2 2 0 9 female 75 plus East NA  
## 3 3 0 10 male East NA  
## 4 4 0.0370 27 female 75 plus East NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers



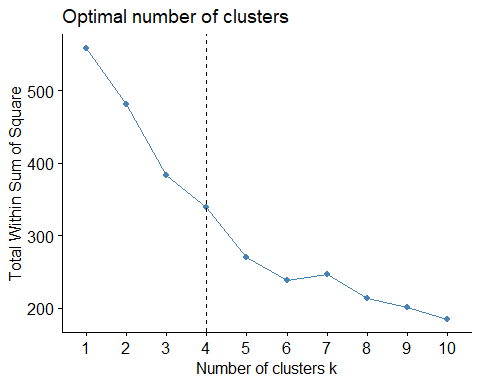
Cluster 4 seems to be the group of interest since it has the highest proportion of individuals who produce more than one form while speaking. The predominant characteristics of this cluster are “female 75 plus” from the “East” area.

## 2. Person Genitive Plural

# Read in the data  
data <- PerGenPlu.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -20.5644 1949.1692 -0.011 0.992  
## Gender.agefemale under 75 0.8405 0.9282 0.905 0.365  
## Gender.agemale 1.2168 0.9027 1.348 0.178  
## AreaPoland 1.4176 1.0134 1.399 0.162  
## AreaWest 0.6522 0.8722 0.748 0.455  
## Length2 17.6867 1949.1690 0.009 0.993  
## Length3 18.0217 1949.1690 0.009 0.993  
## Length4 19.5480 1949.1691 0.010 0.992  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 64.144 on 62 degrees of freedom  
## Residual deviance: 52.349 on 55 degrees of freedom  
## AIC: 68.349  
##   
## Number of Fisher Scoring iterations: 17

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



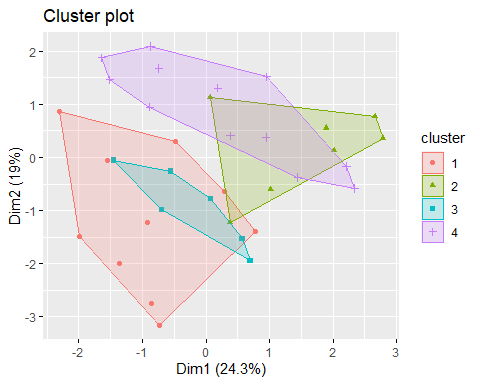
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0.25 16 male Poland NA  
## 2 2 0.4 10 female 75 plus East NA  
## 3 3 0.2 10 male East NA  
## 4 4 0.111 27 female 75 plus East NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

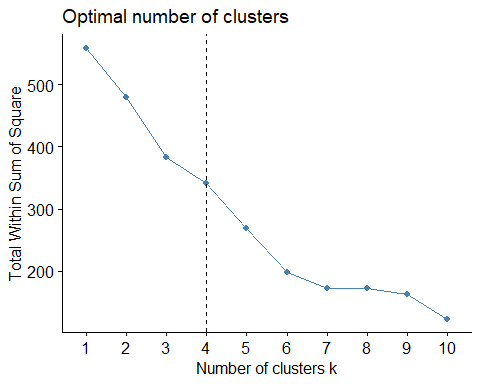
 Cluster 2 stands out as the group with the highest proportion of individuals producing more than one form while speaking. This cluster predominantly comprises older females (75 plus) from the “East” area. Cluster 1, with males from Poland, and Cluster 3, with males from East, also have notable proportions of individuals producing more than one form. Cluster 4, despite being the largest cluster, has the lowest proportion of individuals producing more than one form.

## 3.Person Nominative Singular

# Read in the data  
data <- PerNomSin.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -6.937e+01 1.000e+05 -0.001 0.999  
## Gender.agefemale under 75 -4.541e+01 6.436e+04 -0.001 0.999  
## Gender.agemale -4.589e+01 6.598e+04 -0.001 0.999  
## AreaPoland 4.588e+01 6.468e+04 0.001 0.999  
## AreaWest -8.096e-01 8.047e+04 0.000 1.000  
## Length2 1.094e+00 8.814e+04 0.000 1.000  
## Length3 4.541e+01 8.392e+04 0.001 1.000  
## Length4 4.360e+01 1.402e+05 0.000 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1.027e+01 on 62 degrees of freedom  
## Residual deviance: 2.074e-09 on 55 degrees of freedom  
## AIC: 16  
##   
## Number of Fisher Scoring iterations: 25

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



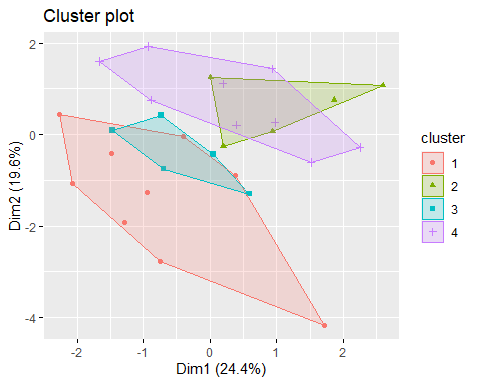
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0.0625 16 male Poland NA  
## 2 2 0 10 female 75 plus East NA  
## 3 3 0 10 male East NA  
## 4 4 0 27 female 75 plus East NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

 Cluster 1 is the only group where a fraction of individuals produce more than one form while speaking. Specifically, males from Poland seem to be the group with this tendency, albeit at a relatively low percentage of 6.25%.

The other clusters (2-4) don’t seem to have any individuals with tot.form equal to 1. This indicates that neither older females from the East nor males from the East show a tendency to produce more than one form in this dataset.

Given that Cluster 4 is the largest cluster and none of its members produce more than one form, it suggests that the “female 75 plus” group from the “East” area predominantly uses a single form when speaking in this dataset.

## 4. Year ADMN

# Read in the data  
data <- YrADMN.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.657e+01 1.566e+05 0 1  
## Gender.agefemale under 75 1.302e-14 1.176e+05 0 1  
## Gender.agemale -1.068e-14 1.162e+05 0 1  
## AreaPoland 1.574e-14 1.272e+05 0 1  
## AreaWest 9.869e-15 1.117e+05 0 1  
## Length2 -1.645e-14 1.461e+05 0 1  
## Length3 -6.413e-15 1.494e+05 0 1  
## Length4 -1.193e-14 1.778e+05 0 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 0.000e+00 on 62 degrees of freedom  
## Residual deviance: 3.655e-10 on 55 degrees of freedom  
## AIC: 16  
##   
## Number of Fisher Scoring iterations: 25

Since non of the individuals used more than one form in this cell, we cant do cluster analysis.

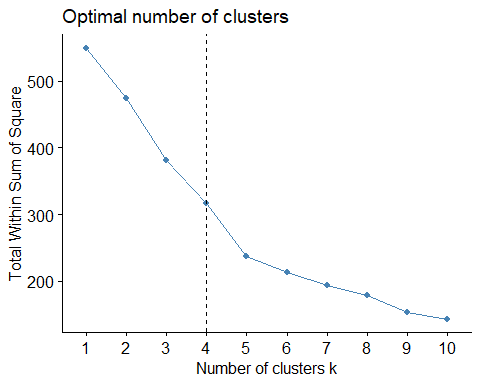
## 5. Year Genitive Plural

# Read in the data  
data <- YrGenPlu.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.5982 1.3384 -1.941 0.0522 .  
## Gender.agefemale under 75 0.5327 0.8522 0.625 0.5319   
## Gender.agemale -0.4849 0.8694 -0.558 0.5770   
## AreaPoland 0.7307 0.9785 0.747 0.4552   
## AreaWest 0.7037 0.8353 0.842 0.3995   
## Length2 0.1372 1.3132 0.104 0.9168   
## Length3 1.6511 1.2263 1.346 0.1782   
## Length4 3.6883 1.4722 2.505 0.0122 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 76.413 on 61 degrees of freedom  
## Residual deviance: 57.918 on 54 degrees of freedom  
## AIC: 73.918  
##   
## Number of Fisher Scoring iterations: 5

The logistic regression model suggests that Length4 has a statistically significant relationship with tot.form.

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



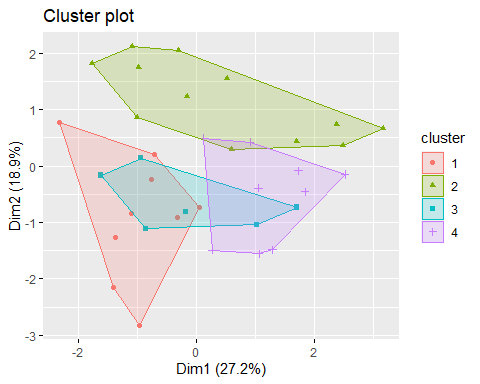
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0.2 15 male Poland NA  
## 2 2 0.304 23 female 75 plus East NA  
## 3 3 0.2 10 male East NA  
## 4 4 0.5 14 female 75 plus West NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

 Cluster 4 stands out as the group with the highest proportion of individuals producing more than one form while speaking. Specifically, older females from the West seem to be the group with this tendency, with half of them producing more than one form.

Cluster 2, comprising older females from the East, also shows a relatively high tendency, with approximately 30.4% of them producing more than one form.

Both Cluster 1 and Cluster 3 have the same proportion of individuals producing more than one form (20%), but they differ in the predominant area.

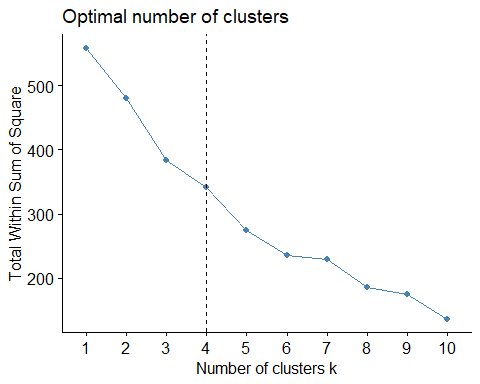
Given this clustering result, the data suggests that older females, whether from the East or West, have a higher likelihood of producing more than one form while speaking compared to males in the dataset

## 6. Year Genitive Singular

# Read in the data  
data <- YrGenSin.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.66332 1.65007 -2.220 0.0264 \*  
## Gender.agefemale under 75 1.28328 1.39236 0.922 0.3567   
## Gender.agemale 2.04208 1.31364 1.555 0.1201   
## AreaPoland -0.13654 1.23133 -0.111 0.9117   
## AreaWest -1.54689 1.40878 -1.098 0.2722   
## Length2 0.06938 1.46632 0.047 0.9623   
## Length3 1.53793 1.34447 1.144 0.2527   
## Length4 1.05417 1.72782 0.610 0.5418   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.953 on 62 degrees of freedom  
## Residual deviance: 37.977 on 55 degrees of freedom  
## AIC: 53.977  
##   
## Number of Fisher Scoring iterations: 6

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



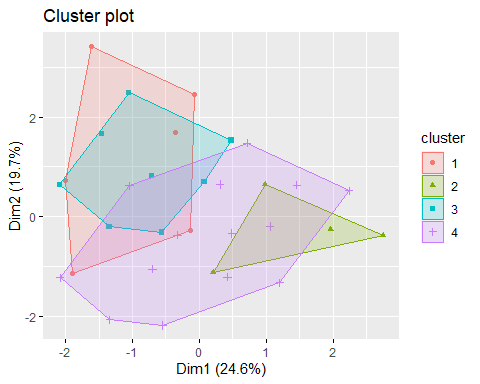
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 1 7 male East NA  
## 2 2 0 9 female 75 plus East NA  
## 3 3 0 15 male East NA  
## 4 4 0 32 female 75 plus West NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

 Cluster 1 stands out as the only group where every individual produces more than one form while speaking. Specifically, males from the East consistently show this tendency.

The other clusters (2-4) show no such tendency. No individual in these clusters produces more than one form

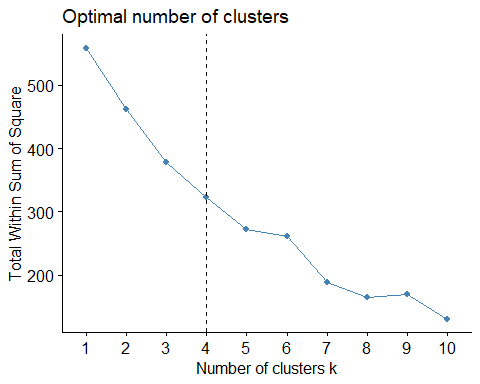
## 7.Year Locatove Singular

# Read in the data  
data <- YrLocSin.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -42.4970 9127.4595 -0.005 0.9963   
## Gender.agefemale under 75 -18.0735 5642.2795 -0.003 0.9974   
## Gender.agemale 2.9589 1.5804 1.872 0.0612 .  
## AreaPoland 18.6615 4605.2757 0.004 0.9968   
## AreaWest 19.7099 4605.2756 0.004 0.9966   
## Length2 19.9496 7880.4793 0.003 0.9980   
## Length3 21.0606 7880.4793 0.003 0.9979   
## Length4 0.9178 10679.2758 0.000 0.9999   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 34.929 on 62 degrees of freedom  
## Residual deviance: 15.017 on 55 degrees of freedom  
## AIC: 31.017  
##   
## Number of Fisher Scoring iterations: 20

The majority of predictors in the model aren’t statistically significant, with the exception of Gender.agemale, which is borderline significant. This suggests that being male might have some influence on the outcome, but the evidence isn’t very strong.

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



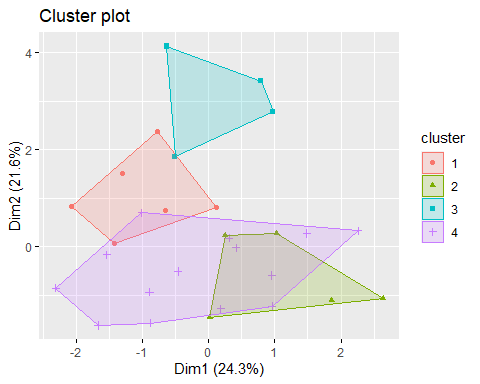
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0 14 male Poland NA  
## 2 2 0 10 female 75 plus East NA  
## 3 3 1 5 male West NA  
## 4 4 0 34 female under 75 East NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

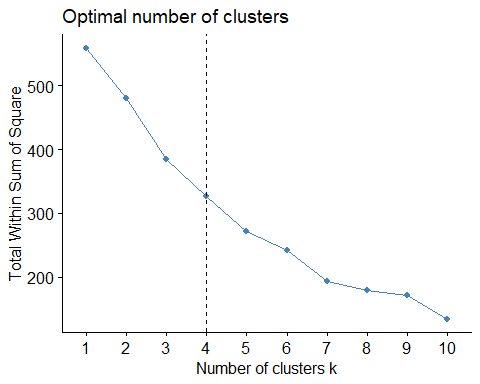
 Cluster 3 stands out as the only cluster where individuals, on average, use forms. This cluster is smaller in size and is characterized by males from the West.

## 8. Year Nominative Singular

# Read in the data  
data <- YrNomSin.wide  
  
# Transform 'tot.form' into binary  
data$tot.form <- ifelse(data$tot.form > 1, 1, 0)  
  
# Fit logistic regression model  
model <- glm(tot.form ~ Gender.age + Area + Length, data=data, family=binomial(link="logit"))  
  
# View the model summary  
summary(model)

##   
## Call:  
## glm(formula = tot.form ~ Gender.age + Area + Length, family = binomial(link = "logit"),   
## data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.153e+02 1.258e+05 -0.001 0.999  
## Gender.agefemale under 75 1.440e-01 7.533e+04 0.000 1.000  
## Gender.agemale 4.564e+01 6.064e+04 0.001 0.999  
## AreaPoland 1.310e+00 7.760e+04 0.000 1.000  
## AreaWest 4.619e+01 6.142e+04 0.001 0.999  
## Length2 4.542e+01 8.467e+04 0.001 1.000  
## Length3 -6.659e-02 9.563e+04 0.000 1.000  
## Length4 9.508e-02 1.067e+05 0.000 1.000  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1.0270e+01 on 62 degrees of freedom  
## Residual deviance: 2.1263e-09 on 55 degrees of freedom  
## AIC: 16  
##   
## Number of Fisher Scoring iterations: 25

# Data preparation  
# Convert categorical variables to binary format  
data <- data %>% select(Gender.age, Area, Length, tot.form) # Select relevant columns  
data\_dummies <- model.matrix(~ . - 1, data) # Create dummies, dropping the first level of each factor  
  
# Standardize the data  
data\_standardized <- scale(data\_dummies)  
  
# Determine the number of clusters using Elbow method  
fviz\_nbclust(data\_standardized, kmeans, method = "wss") + geom\_vline(xintercept = 4, linetype = 2)



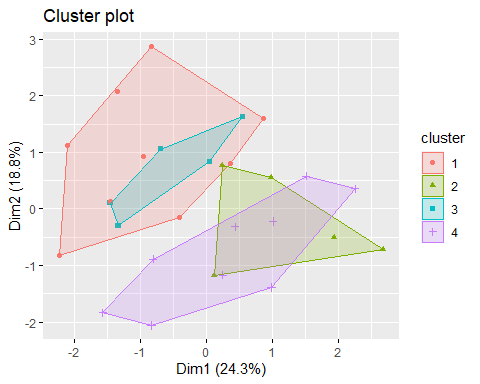
# Assuming you decide on 'k' clusters based on the Elbow method, run k-means clustering  
set.seed(123) # Setting seed for reproducibility  
kclusters <- kmeans(data\_standardized, centers = 4) # replace 'k' with the optimal number of clusters  
  
# Attach cluster assignment to data  
data$cluster <- kclusters$cluster  
  
# Analyze clusters  
cluster\_summary <- data %>%   
 group\_by(cluster) %>%   
 summarise(  
 avg\_tot\_form = mean(tot.form),  
 count = n(),  
 predominant\_gender\_age = names(sort(table(Gender.age), decreasing = TRUE)[1]),  
 predominant\_area = names(sort(table(Area), decreasing = TRUE)[1]),  
 avg\_length = mean(Length)  
 )

## Warning: There were 4 warnings in `summarise()`.  
## The first warning was:  
## ℹ In argument: `avg\_length = mean(Length)`.  
## ℹ In group 1: `cluster = 1`.  
## Caused by warning in `mean.default()`:  
## ! argument is not numeric or logical: returning NA  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

print(cluster\_summary)

## # A tibble: 4 × 6  
## cluster avg\_tot\_form count predominant\_gender\_age predominant\_area avg\_length  
## <int> <dbl> <int> <chr> <chr> <dbl>  
## 1 1 0 16 male Poland NA  
## 2 2 0 10 female 75 plus East NA  
## 3 3 0.1 10 male East NA  
## 4 4 0 27 female 75 plus East NA

fviz\_cluster(kclusters, data = data\_standardized,   
 geom = "point", # Only show data points  
 show.clust.cent = FALSE) # Don't show cluster centers

 Clusters 1 and 3 are dominated by males, but they differ in their primary regions (Poland vs. East). Clusters 2 and 4 are dominated by females aged 75 and above from the East. Cluster 4 is notably larger than Cluster 2. The characteristic or centroid value seems to be relatively consistent across the clusters, with only Cluster 3 showing a minor difference.