

MixAll_updated

PMS

2023-05-29

##	CSDUID	DBUID	DBRPLAMX	DBRPLAMY	PRUID					
## 1	1001519	10010202002	8979444	2148775	10					
## 2	1001519	10010203001	8979186	2149065	10					
## 3	1001519	10010204001	8979382	2148616	10					
## 4	1001519	10010204002	8979490	2148537	10					
## 5	1001519	10010204004	8979255	2148525	10					
##						PRNAME	CDUID		CDNAME	
## 1	Newfoundland and Labrador	/	Terre-Neuve-et-Labrador			1001	Division No.		1	
## 2	Newfoundland and Labrador	/	Terre-Neuve-et-Labrador			1001	Division No.		1	
## 3	Newfoundland and Labrador	/	Terre-Neuve-et-Labrador			1001	Division No.		1	
## 4	Newfoundland and Labrador	/	Terre-Neuve-et-Labrador			1001	Division No.		1	
## 5	Newfoundland and Labrador	/	Terre-Neuve-et-Labrador			1001	Division No.		1	
##	CDTYPE	CCSUID	CCSNAME	CSDNAME.x	CSDTYPE	ERUID		ERNAME	FEDUID	
## 1	CDR	1001519	St. John's	St. John's	CY	1010	Avalon Peninsula		10007	
## 2	CDR	1001519	St. John's	St. John's	CY	1010	Avalon Peninsula		10007	
## 3	CDR	1001519	St. John's	St. John's	CY	1010	Avalon Peninsula		10007	
## 4	CDR	1001519	St. John's	St. John's	CY	1010	Avalon Peninsula		10007	
## 5	CDR	1001519	St. John's	St. John's	CY	1010	Avalon Peninsula		10007	
##								FEDNAME	SACCODE	SACTYPE
## 1	St. John's	South--Mount Pearl	/	St. John's-Sud--Mount Pearl				001		1
## 2	St. John's	South--Mount Pearl	/	St. John's-Sud--Mount Pearl				001		1
## 3	St. John's	South--Mount Pearl	/	St. John's-Sud--Mount Pearl				001		1
## 4	St. John's	South--Mount Pearl	/	St. John's-Sud--Mount Pearl				001		1
## 5	St. John's	South--Mount Pearl	/	St. John's-Sud--Mount Pearl				001		1
##	CMAUID	CMAPUID	CMAName	CMAType	CTUID	CTName	ADAUID		DAUID	
## 1	001	10001	St. John's	B	0010003.01	0003.01	10010015		10010202	
## 2	001	10001	St. John's	B	0010003.01	0003.01	10010015		10010203	
## 3	001	10001	St. John's	B	0010003.01	0003.01	10010015		10010204	
## 4	001	10001	St. John's	B	0010003.01	0003.01	10010015		10010204	
## 5	001	10001	St. John's	B	0010003.01	0003.01	10010015		10010204	
##	PMS_DBPOP	PMS_DAUID	PMS_DAPOP	PMS_CSDUID	PMS_CSDNAME	PMS_CSDTYPE		PMS_CSDPOP		
## 1	116	10010202	444	1001519	St. John's	CY		108,860		
## 2	350	10010203	350	1001519	St. John's	CY		108,860		
## 3	66	10010204	541	1001519	St. John's	CY		108,860		
## 4	0	10010204	541	1001519	St. John's	CY		108,860		
## 5	177	10010204	541	1001519	St. John's	CY		108,860		
##	PMS_CMAUID	PMS_CMAPUID	PMS_CMAName	PMS_CMAType	PMS_CMAPOP		PMS_PRUID			
## 1	1	10001	St. John's	B	205,955		10			
## 2	1	10001	St. John's	B	205,955		10			
## 3	1	10001	St. John's	B	205,955		10			
## 4	1	10001	St. John's	B	205,955		10			
## 5	1	10001	St. John's	B	205,955		10			

```

##                                PMS_PRNAME PMS_PRPOP  PMS_lon
## 1 Newfoundland and Labrador / Terre-Neuve-et-Labrador  519,716 -52.7463
## 2 Newfoundland and Labrador / Terre-Neuve-et-Labrador  519,716 -52.7469
## 3 Newfoundland and Labrador / Terre-Neuve-et-Labrador  519,716 -52.7482
## 4 Newfoundland and Labrador / Terre-Neuve-et-Labrador  519,716 -52.7476
## 5 Newfoundland and Labrador / Terre-Neuve-et-Labrador  519,716 -52.7502
##  PMS_lat PMS_in_db_emp PMS_prox_idx_emp PMS_in_db_pharma PMS_prox_idx_pharma
## 1 47.5388           1           0.0218           0           0.0184
## 2 47.5423           1           0.0232           0           0.0068
## 3 47.5380           1           0.0218           0           0.0316
## 4 47.5368           0           0.0209           0           0.0316
## 5 47.5379           1           0.0205           0           0.0316
##  PMS_in_db_childcare PMS_prox_idx_childcare PMS_in_db_health
## 1           0           0.0155           0
## 2           0           0.0158           1
## 3           0           0.0126           0
## 4           0           0.0124           0
## 5           0           0.0056           0
##  PMS_prox_idx_health PMS_in_db_grocery PMS_prox_idx_grocery PMS_in_db_educpri
## 1           0.0076           0           NA           0
## 2           0.0095           0           0.0346           0
## 3           0.0073           0           NA           0
## 4           0.0073           0           NA           0
## 5           0.0033           0           NA           0
##  PMS_prox_idx_educpri PMS_in_db_educsec PMS_prox_idx_educsec PMS_in_db_lib
## 1           NA           0           NA           0
## 2           NA           0           NA           0
## 3           NA           0           NA           0
## 4           NA           0           NA           0
## 5           NA           0           NA           0
##  PMS_prox_idx_lib PMS_in_db_parks PMS_prox_idx_parks PMS_in_db_transit
## 1           NA           0           NA           1
## 2           NA           0           NA           1
## 3           NA           0           NA           1
## 4           NA           0           NA           0
## 5           NA           0           NA           0
##  PMS_prox_idx_transit PMS_transit_na PMS_amenity_dense PMS_suppressed
## 1           0.0155           0           0           0
## 2           0.0069           0           0           0
## 3           0.0117           0           0           0
## 4           0.0184           0           0           0
## 5           0.0091           0           0           0
##  CSDNAME.y PROVINCE BUILD_AREA AVG_LAT AVG_LONG NUMBER_BUILD CSD_AREA IOR_X
## 1    <NA>    <NA>      NA      NA      NA      NA      NA      89
## 2    <NA>    <NA>      NA      NA      NA      NA      NA      89
## 3    <NA>    <NA>      NA      NA      NA      NA      NA      89
## 4    <NA>    <NA>      NA      NA      NA      NA      NA      89
## 5    <NA>    <NA>      NA      NA      NA      NA      NA      89
##  IOR_Pruid IOR_CSDname IOR_CSDtype IOR_Index_of_remoteness IOR_CSDpop2021
## 1       10 St. John's      CY      0.247038      110525
## 2       10 St. John's      CY      0.247038      110525
## 3       10 St. John's      CY      0.247038      110525
## 4       10 St. John's      CY      0.247038      110525
## 5       10 St. John's      CY      0.247038      110525

```

```
##          IOR_DGUID
## 1 2021A00051001519
## 2 2021A00051001519
## 3 2021A00051001519
## 4 2021A00051001519
## 5 2021A00051001519
```

subsampling data

```
# covert populations to numeric from factor
master$PMS_DBPOP = as.numeric(gsub("[^0-9.-]", "", as.character(master$PMS_DBPOP)))
master$PMS_CSDPOP = as.numeric(gsub("[^0-9.-]", "", as.character(master$PMS_CSDPOP)))

# subsampling
perc = 3 #percentage of data to subsample
subsample = (nrow(master)/100)*perc
pms_subsample = master[sample(nrow(master), subsample),]

ammenities <- c("PMS_prox_idx_emp", "PMS_prox_idx_pharma", "PMS_prox_idx_childcare", "PMS_prox_idx_heal

pms_prox <- pms_subsample[, ammenities]
```

subsampling data

```
# results_df <- data.frame()
#
# for (amenity in ammenities) {
#
#   for (num_clusters in 2:8) {
#     # each amenity
#     pms_amen <- pms_prox[, amenity]
#     pms_amen <- na.omit(pms_amen)
#     pms_amen_log <- log(pms_amen + 0.0001)
#
#     # cluster data
#     model <- Mclust(pms_amen_log, G = num_clusters)
#     clus_labels = model$classification
#     clus_data <- model$data
#
#     # calculate metrics
#     dunn_stats <- cluster.stats(dist(clus_data), clus_labels)
#     calinski_harabasz <- calinhara(clus_data, clus_labels)
#     xie_beni <- intCriteria(as.matrix(clus_data), as.integer(clus_labels), 'Xie_Beni')$xie_beni
#     davies_bouldin <- index.DB(clus_data, clus_labels)
#     sil_coef = intCriteria(as.matrix(clus_data), as.integer(clus_labels), 'Silhouette')$silhouette
#
#     # df containing each amenity, # of clusters, and metric score
#     amenity_df <- data.frame(
#       amenity = amenity,
#       num_clusters = num_clusters,
```

```

#       dunn = dunn_stats$dunn,
#       calinski_harabasz = calinski_harabasz,
#       xie_beni = xie_beni,
#       davies_bouldin = davies_bouldin$DB,
#       sil_coef = sil_coef
#     )
#
#     results_df <- rbind(results_df, amenity_df)
#   }
# }
#
# head(results_df, 7)

```

Cutoff function

number of clusters function

getting cutoffs and metrics for optimal cluster

```

sil_mixall <- c()
dunn_mixall <- c()
xei_beni_mixall <- c()
calinski_harabasz_mixall <- c()
davies_bouldin_mixall <- c()
num_of_clust_mixall <- c()

cut_offs_mixall <- list()

for (amenity in ammenities) {

  # select the best number of clusters for current amenity
  # num_clusters <- optimal_num_clusters(amenity, results_df)

  pms_amen <- pms_prox[, amenity]
  pms_amen <- na.omit(pms_amen)
  pms_amen_log <- log(pms_amen + 0.0001)

  # cluster data
  model <- Mclust(pms_amen_log)

  clus_labels = model$classification
  clus_data <- model$data
  num_clusters <- model$G

  # compute metrics and cutoffs for storing later on
  dunn_stats <- cluster.stats(dist(clus_data), clus_labels)
  calinski_harabasz <- calinhara(clus_data, clus_labels)
  xie_beni <- intCriteria(as.matrix(clus_data), as.integer(clus_labels), 'Xie_Beni')$xie_beni
  davies_bouldin <- index.DB(clus_data, clus_labels)
  sil_coef <- intCriteria(as.matrix(clus_data), as.integer(clus_labels), 'Silhouette')$silhouette
  cutoffs <- compute_cutoffs(clus_data, clus_labels)
}

```

```

# store the metrics
sil_mixall <- c(sil_mixall, sil_coef)
dunn_mixall <- c(dunn_mixall, dunn_stats$dunn)
xei_beni_mixall <- c(xei_beni_mixall, xie_beni)
calinski_harabasz_mixall <- c(calinski_harabasz_mixall, calinski_harabasz)
davies_bouldin_mixall <- c(davies_bouldin_mixall, davies_bouldin$DB)
num_of_clust_mixall <- c(num_of_clust_mixall, num_clusters)
cut_offs_mixall[[amenity]] <- cutoffs
}

```

```
sil_mixall
```

```
## [1] 0.5867320 0.4837194 0.6027205 0.5222099 0.5909881 0.4577264 0.4784640
## [8] 0.4892339 0.4577640 0.5762659
```

```
dunn_mixall
```

```
## [1] 0.0012634720 0.0001982587 0.0003168529 0.0023371238 0.0011548845
## [6] 0.0004263013 0.0004016697 0.0016705470 0.0001100838 0.0024851275
```

```
xei_beni_mixall
```

```
## [1] 21750.3878 95888.8433 152342.6006 3189.4805 8495.1027 53343.1775
## [7] 66257.9774 704.2013 67763.6691 5389.6353
```

```
calinski_harabasz_mixall
```

```
## [1] 98538.764 4928.255 9950.952 23477.133 1959.630 18423.626 7935.748
## [8] 4323.034 17243.941 11501.814
```

```
davies_bouldin_mixall
```

```
## [1] 0.5575865 25.1732969 0.6417944 0.6392660 0.6932940 0.6476328
## [7] 0.5943470 0.6776949 0.9642396 0.6424359
```

```
num_of_clust_mixall
```

```
## [1] 9 7 3 4 3 7 8 7 8 3
```

```
cut_offs_mixall
```

```
## $PMS_prox_idx_emp
## [1] -8.863767 -7.509742 -6.608337 -5.669486 -4.750213 -3.875212 -2.957474
## [8] -1.813699
##
## $PMS_prox_idx_pharma
## [1] -5.033239 -4.519035 -4.003594 -3.402238 -2.890984 -4.468784
##
```

```

## $PMS_prox_idx_childcare
## [1] -6.227267 -2.703809
##
## $PMS_prox_idx_health
## [1] -7.967887 -5.640907 -4.661755
##
## $PMS_prox_idx_grocery
## [1] -4.440518 -4.917167
##
## $PMS_prox_idx_educpri
## [1] -3.746544 -3.626533 -3.112269 -2.406280 -2.030651 -1.686859
##
## $PMS_prox_idx_educsec
## [1] -3.356711 -3.361157 -3.124703 -2.782256 -2.290166 -1.941072 -2.458470
##
## $PMS_prox_idx_lib
## [1] -3.017986 -2.920632 -2.683127 -2.376788 -2.150301 -3.174928
##
## $PMS_prox_idx_parks
## [1] -2.493186 -4.132051 -3.424979 -3.069379 -2.773389 -2.470412 -2.103325
##
## $PMS_prox_idx_transit
## [1] -6.768939 -4.443911

```