College Scorecard: Cluster Analysis

Michael L. Thompson

September 4, 2017

Contents

Introduction

This is an exploratory analysis of the U.S. Dept. of Education College Scorecard database. My intent is to investigate patterns amongst the colleges as visualized using t-distributed Stochastic Neighbor Embedding (t-SNE)[1] using the R package **Rtsne**[2]. This method projects the high-dimensional data into two dimensions. From there, I can apply hierarchical clustering to identify clusters in the new 2-D space.

Setup

First, load packages from the local library. . . .

Note: The package **GraphAlignment** was downloaded and installed from BioConductor using the following R commands:

[1] L.J.P. van der Maaten. **Accelerating t-SNE using Tree-Based Algorithms.** *Journal of Machine Learning Research* 15(Oct):3221-3245, 2014. PDF [Supplemental material]

[2] Jesse H. Krijthe (2015). **Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation**, URL: https://github.com/jkrijthe/Rtsne

```
source("http://bioconductor.org/biocLite.R")
biocLite("GraphAlignment")
```

Prepare Data

We read in the College Scorecard dataset and convert columns into Bayes factors, which accentuate differences amongst the colleges. Colleges having a disproportionately high number of students with a certain attribute – say, an SAT in excess of 1400 – will have highly positive Bayes factors for that attribute.

I strip out a lot of the variables that define the student body demographics. The idea is that I'd like to identify structure in the "outcome" variables – things like academic disciplines, completion rates, future earnings, credit default rates, etc. – and then later check if this structure is correlated to demographics – things like geographic location, campus setting, student ethnicity, etc.

```
glmdata_all <- DataSpec$studentBF %>%
  dplyr::select(
    c(-1, -(3:8)), -matches('_(WHITE|BLACK|ASIAN|OTHER|HISP|NRA|AIAN|UNKN)|2MOR|UNKN|NHPI|AIAN|BF_male|
    -matches('Challenge|_DEP_STAT_|notvet|le24y|OUTOFSTATE|prior|(^BF_[gl][et].+[0-9]+K$)|locale|FarWes
  ) %>%
  select_if( .predicate = function(x) any(x != x[[1]]) ) %>%
  filter( complete.cases(.) )
tsne_mat_all <- glmdata_all %>% select(-College) %>% as.matrix() %>% scale()
```

Perform t-Distributed Stochastic Neighbor Embedding (t-SNE)

Now, I'll map the data into a 2-D space using t-SNE. Hopefully, it will be easy to see clusters of colleges.

It takes a bit of trial and error (short of doing a formal hyperparameter optimization) to arrive at hyperparameters capable of generating discernible structure in a 2-D scatterplot.

```
set.seed( 173 )
tsne_all <- Rtsne( tsne_mat_all, perplexity = 10, initial_dims = 50, theta = 0.5, max_iter = 2000 )
```

Rotate Coordinates

Now, I'll rotate the coordinates so that high-prestige colleges appear at high `Y2` coordinates. This will put most of the Ivy League colleges in the top-center of the plot.

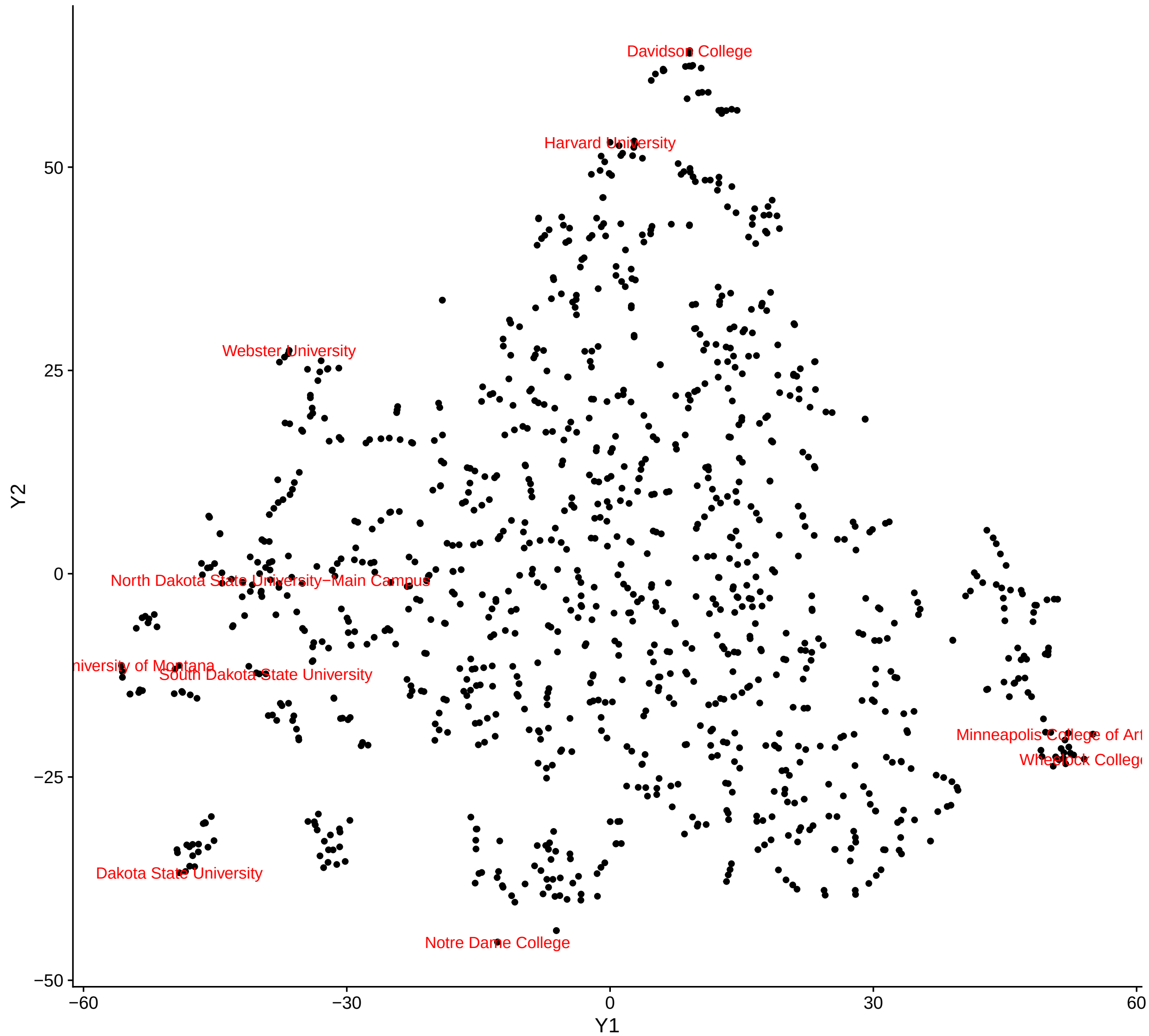
```
# Rotate coordinates so that high-prestige colleges appear at high Y2 coordinates,
# i.e. in the top center of the plot.
i_harvard <- grep( 'Harvard', glmdata_all$College )
harvard_coord <- tsne_all$Y[i_harvard,]
harvard_angle <- atan(harvard_coord[2]/harvard_coord[1])
rotate_angle  <- pi/2 - harvard_angle
rotation_matrix <- matrix(
  c(cos(rotate_angle),sin(rotate_angle),-sin(rotate_angle),cos(rotate_angle)),
  2,2, byrow = TRUE
)
tsne_all$Y %<>% { (.) %*% rotation_matrix }
if( abs(tsne_all$Y[i_harvard,2]) < abs(tsne_all$Y[i_harvard,1]) ){
  tmp <- tsne_all$Y[,1]
  tsne_all$Y[,1] <- tsne_all$Y[,2]
  tsne_all$Y[,2] <- tmp
```

```
}
if( tsne_all$Y[i_harvard,2] < 0 ){
  tsne_all$Y[,2] <- -tsne_all$Y[,2]
}
```

I'll highlight colleges at the minimum and maximum of each of the coordinate axes and diagonals. This is done by projecting each college's coordinates onto vectors pointing into those 4 direction vectors – up, right, top-right, top-left – and finding the colleges that are at the maximum positive and negative points along those vectors.

The names of those colleges at the extremes are added along with "Harvard" as names to be highlighted in the 2-D scatterplot.

```
# Project each college's coordinates along the 4 direction vectors.
prj <- tsne_all$Y %*% matrix(c(1,0,0,1,1,1,-1,1),nrow=2,ncol=4)

# Identify the colleges to be highlighted as Harvard and those at the min and max of the direction vect
highlights <- union(
  'Harvard',
  as.character(glmdata_all$College)[c(apply(prj,2,function(x) c(which.min(x),which.max(x))))]
) %>%
  paste(collapse="|")

# Plot the 2D scatterplot with highlighted colleges labeled by the college name.
tsne_all$Y %>%
  as_tibble() %>%
  setNames(c('Y1','Y2')) %>%
  mutate( College = glmdata_all$College) %>%
  {
    ggplot(.,aes(x=Y1,y=Y2)) +
      geom_point() +
      geom_text(
        data    = (.) %>% filter( grepl(highlights,College) ),
        mapping = aes( label = College ),
        color   = 'red',
        size    = 4
      )
  } %>%
  print()
```



Find Underlying Factors Driving 2-D Structure

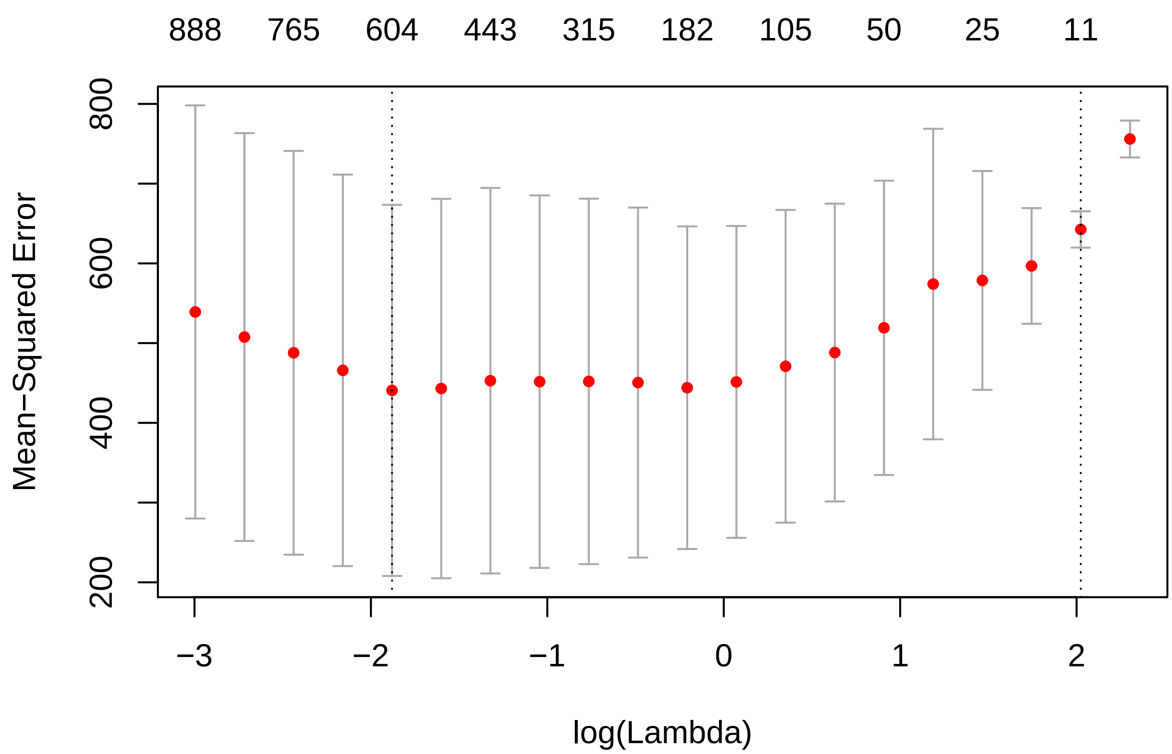
Using R package **glmnet**[3], I perform regularization (variable selection) in modeling of the 2-D t-SNE coordinates as responses vs. the original college Bayes factor features from which the t-SNE coordinates were found. This way we'll have a linear model showing which features contributed to which coordinate. As such, we'll have the basis for plotting a biplot of colleges overlayed on feature dimensions in 2-D, analogous to a PCA biplot.

```
mmat <- model.matrix( ~ .:. - 1, as.data.frame(tsne_mat_all))
# b <- eigen(cor(mmat))
# mmat <- mmat[,apply(b$vectors[,1:200],2,function(x) which.max(abs(x))) %>% unique() %>% sort()]
```

```
set.seed( 2393 )
tsne_glmnet_all <- cv.glmnet(
  x       = mmat,
```

[3] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *Journal of Statistical Software*, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

```
  y      = tsne_all$Y,
  family = 'mgaussian',
  lambda = exp(seq(log(0.05),log(10),length.out = 20))
)
plot( tsne_glmnet_all )
```



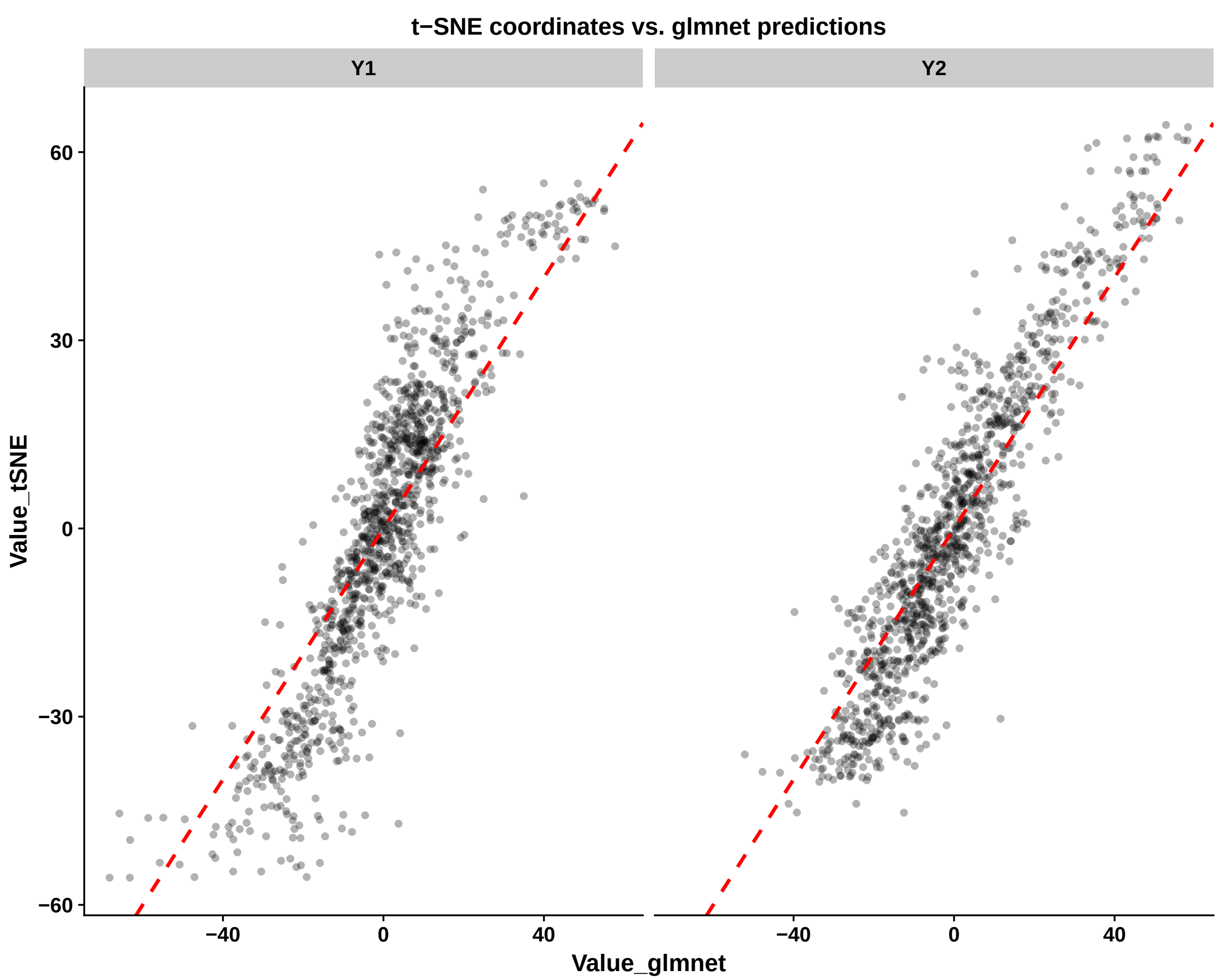
Check the Predictions

It can be tricky to find a subset of features and their interactions that both describe the t-SNE coordinates well *and* do not suffer from extreme collinearity, which can make the validation error at low `lambda` explode when applying function `cv.glmnet()`.

Judging from the cross-validation curve above and the observed vs. predicted plots below, it looks like we've got a decent model.

```
# Get glmnet predictions of the t-SNE coordinates, combine them with the original t-SNE coordinates,
# and plot the originals vs. predictions.
lambda <- tsne_glmnet_all %$% { exp( mean(log(c(lambda.min,lambda.1se))) ) } # mid lambda
lambda <- 1.87654485

stack_coords <- function( coord_matrix ){
  coord_matrix %>%
    as_data_frame() %>%
    setNames( c( "Y1", "Y2" ) ) %>%
    mutate( rowid = 1:nrow(.) ) %>%
    gather( key = Coordinate, value = Value , -rowid )
```

```r
}

# Plot t-SNE coords. vs. glmnet prediction of t-SNE coords.
tsne_glmnet_all %>%
  predict( newx = mmat, s = lambda ) %>%
  drop() %>%
  stack_coords() %>%
  left_join(
    y      = tsne_all$Y %>% stack_coords(),
    by     = c('Coordinate','rowid'),
    suffix = c( "_glmnet","_tSNE" )
  )  %>%
  {
    ggplot(., aes(x = Value_glmnet, y = Value_tSNE ) ) +
      geom_point( alpha = 0.3 ) +
      geom_abline( intercept = 0, slope = 1, color = 'red', linetype = 2, size = 1 ) +
      facet_wrap( ~ Coordinate ) +
      ggtitle( "t-SNE coordinates vs. glmnet predictions" ) +
      theme( text = element_text( face = 'bold' ) )
  } %>%
  print()
```



Visualize the Colleges in 2-D

Analogous to PCA, which has component (i.e., factor) loading vectors defining the basis vectors (dimensions) of the space and has a scores matrix defining the position of the items in the space, we'll use the 2-D t-SNE coordinates as our "factors" and as such the **glmnet** coefficients are the factor "loadings" projecting the raw Bayes factor features of the colleges (our "items") into the 2-D space. Therefore the t-SNE coordinates of the colleges serve as the "scores" matrix.

Get Factor "Loadings" from glmnet Coefficients

Now the **glmnet** model coefficients will serve as the basis vectors (dimensions) of the biplot, since the model is simply a linear combo of the feature coefficients and each college's values for the respective features.

```
# Get the sparse matrix of coefficients and strip down to only the non-zero coefficients.
tsne_glmnet_coef_all <- tsne_glmnet_all %>% coef( s = lambda )

tsne_coef_df_all <-
  tsne_glmnet_coef_all$y1 %>%
  as.matrix() %>%
  as.data.frame() %>%
  as_tibble() %>%
  rownames_to_column() %>%
  setNames(c("Coefficient","Y1")) %>%
  full_join(
    tsne_glmnet_coef_all$y2 %>%
      as.matrix() %>%
      as.data.frame() %>%
      as_tibble() %>%
      rownames_to_column() %>%
      setNames(c("Coefficient","Y2")),
    by = "Coefficient"
  ) %>%
  filter( abs(Y1) > 1.0E-9 | abs(Y2) > 1.0E-9 ) %>% slice(-1) %>%
  mutate(
    # Flip direction of interactions
    Y1 = ifelse( grepl(':',Coefficient), -Y1 , Y1 ),
    Y2 = ifelse( grepl(':',Coefficient), -Y2 , Y2 )
  )

tsne_coef_df_all %>%
  mutate(mag = sqrt(Y1^2+Y2^2)) %>%
  arrange(desc(mag)) %>%
  mutate_at(funs(round(.,1)),.vars=vars(-Coefficient)) %>%
  print(n = 30)
```

```
## # A tibble: 73 x 4
##                           Coefficient    Y1    Y2   mag
##                                 <chr> <dbl> <dbl> <dbl>
##  1                     BF_discBreadth  -5.3   0.8   5.4
##  2                BF_ForeignLanguages  -2.5   3.9   4.6
##  3             BF_ScienceTechnologies  -3.4  -2.3   4.1
##  4          BF_AgricultureAgriculture  -2.9  -1.0   3.1
##  5                       BF_SAT_gt1400   0.5   2.9   3.0
##  6                     BF_fsend_5_2005   1.3   2.7   2.9
```

```
## 7                                 BF_CDR3est  -0.5  -2.9   2.9
## 8                         BF_PhysicalSciences  -1.8   2.2   2.8
## 9                           BF_pell_ever_2005  -1.1  -2.3   2.5
## 10                        BF_PersonalCulinary  -2.1  -1.3   2.5
## 11                                 BF_veteran  -1.8  -1.5   2.3
## 12              BF_CommunicationsTechnologies  -2.0   0.6   2.1
## 13       BF_SAT_gt800le1000:BF_MechanicRepair   1.6  -1.0   1.9
## 14                              BF_AreaEthnic  -1.0   1.2   1.6
## 15          BF_Education:BF_TheologyReligious  -1.4  -0.7   1.5
## 16                     BF_PhilosophyReligious  -0.5   1.4   1.5
## 17                  BF_EngineeringTechnologies  -1.3  -0.6   1.4
## 18                        BF_VisualPerforming  -1.4   0.1   1.4
## 19                   BF_MathematicsStatistics  -1.1   0.7   1.3
## 20                          BF_MechanicRepair  -1.3   0.0   1.3
## 21                     BF_ComputerInformation  -1.2   0.4   1.2
## 22                               BF_SAT_le800   0.0  -1.2   1.2
## 23                     BF_ArchitectureRelated  -1.1   0.1   1.2
## 24                        BF_HomelandSecurity  -0.8  -0.7   1.0
## 25                       BF_HealthProfessions  -0.9  -0.4   1.0
## 26     BF_Engineering:BF_PhilosophyReligious   0.9   0.3   1.0
## 27              BF_veteran:BF_PhysicalSciences   0.5   0.8   0.9
## 28                              BF_gt24yrsold  -0.3  -0.9   0.9
## 29 BF_CommunicationsTechnologies:BF_Education   0.8  -0.3   0.9
## 30                          BF_FamilyConsumer  -0.8  -0.3   0.9
## # ... with 43 more rows
```

```
# Designate coeffcient vectors in the top 10% in magnitude as "key terms".
# And sort from largest magnitude to smallest.
key_terms <- tsne_coef_df_all %>%
  mutate(mag= sqrt(Y1^2+Y2^2)) %>%
  filter(abs(mag)>quantile(abs(mag),0.9)) %>%
  arrange(desc(mag)) %$% Coefficient %>% setdiff("(Intercept)")
```

Get "Scores" Matrix from t-SNE Coordinates

The college 2-D t-SNE coordinates are the "scores" matrix, which we collect into a `data_frame` with ancillary info for labeling and coloring in our 2-D scatterplots.

```
# For preliminary coloring in plot, divide colleges into categories that
# capture the 10, 25, 75 and 90 percentiles along the Y2 axis, which has bee
# rotated to point towards Ivy League colleges (specifically towards Harvard U.).
categories <- {
  mmat[,key_terms] %*%
    (tsne_coef_df_all %>% filter(Coefficient %in% key_terms) %$% Y2)
} %>%
  sapply(
    function(x,q){ length(q) - sum(x>q) + 1 },
    q=quantile(.,c(0.1,0.25,0.75,0.9))
  ) %>%
  factor()

# Abbreviate names so they don't clutter the plots so much.
shorten_names <- function( df ){
  college_names <- df %$%
```

```r
    College %>%
    { gsub('^[0-9_]+','',. ) } %>%
    { gsub('The Univer.+ of Texas at ','U.T. ',.) } %>%
    { gsub('Advancement of Science','Adv.Sci',.) } %>%
    { gsub('Northwestern University','NU',.) } %>%
    { gsub('University of Notre Dame','Notre Dame U.',.) } %>%
    { gsub('Cornell College','Cornell C',.) } %>%
    { gsub('Cornell University','Cornell U',.) } %>%
    { gsub('California','Cal',. ) } %>%
    { gsub('Mass.+Inst.+Tech.+','MIT',. ) } %>%
    { gsub('(Mass|Penn|Wash)[^ ]+ *','\\1',.) } %>%
    { gsub('Polytechnic','Poly',. ) } %>%
    { gsub('Institute of Tech[^ ]+','IT',. ) } %>%
    { gsub('Tech.+Inst.+','Tech',. ) } %>%
    { gsub('State','St',. ) } %>%
    { gsub('University','U',. ) } %>%
    { gsub('(U of )|( U$)','',. ) } %>%
    { gsub('College','Col',. ) } %>%
    { gsub('New York','NY',.)} %>%
    { gsub('International','Intl',.) } %>%
    { gsub('North[^ ]+','N',.)} %>%
    { gsub('South[^ ]+','S',.)} %>%
    { gsub('West[^ ]+','W',.)} %>%
    { gsub('East[^ ]+','E',.)} %>%
    { gsub(' U-','-',.)} %>%
    { gsub('-Penn St ','',.)} %>%
    { gsub(' Col *$','',.)} %>%
    { gsub('-(Main)* Campus','',.)} %>%
    { gsub('^PennSt([^-]+)$','Penn St-\\1',.)} %>%
    { gsub(' and ','&',.)} %>%
    { gsub('Agricultural & Mechanical','A&M',.)}

  st_abb <- state.abb %>% setNames( state.name )
  for( st_nm in names(st_abb) ){
    college_names %<>% { gsub(st_nm,st_abb[st_nm],.) }
  }
  return( college_names )
}
college_names <- shorten_names( glmdata_all )
college_names_student <- shorten_names( DataSpec$student)

# Collect all colleges and their t-SNE coords, Bayes factor for high-income, and category designators i
tsne_df_all <- tsne_all$Y %>%
  as_tibble() %>%
  setNames(c("Y1","Y2")) %>%
  mutate(
    College = college_names,
    category = categories,
    BF_Income_gt110K = glmdata_all %$% {10.0^BF_p_gt110K}
  ) %>%
  dplyr::select( College, category, BF_Income_gt110K, everything() ) %>%
  mutate_at(funs(scale(.)),.vars=vars(Y1,Y2))
```

Show Biplot for Structure Interpretation

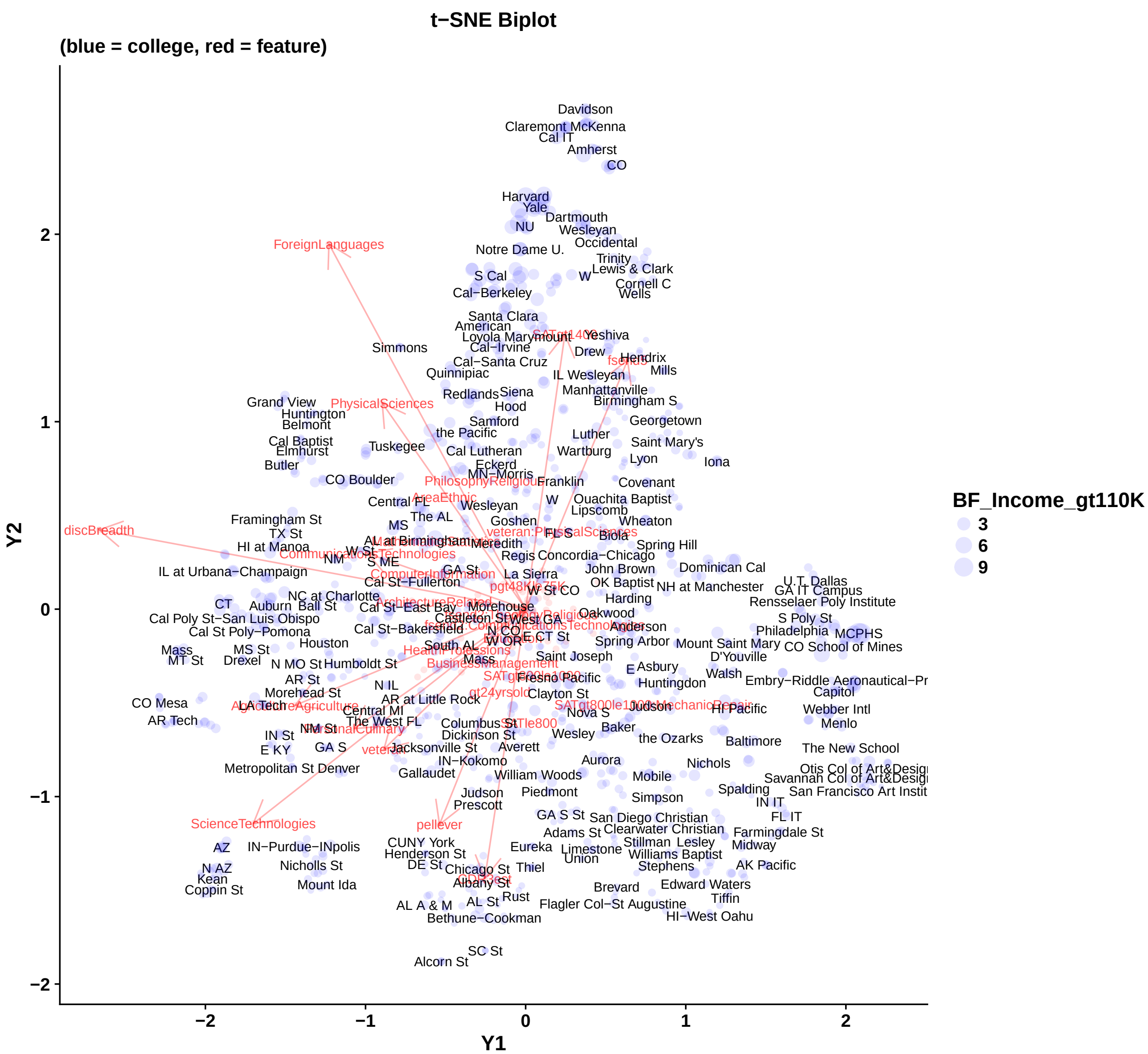
As with a PCA biplot, we can overlay the feature dimensions on the college scatterplot in the 2-D t-SNE coordinate space.
This allows us to more easily interpret the structure we're seeing.

However, some of the interaction terms, in particular, are tricky to interpret because they have a positive value for a college if both of the features in the product making up the interaction have the same sign. So it could be that the college has a disproportionately higher *or* lower number of students having the attributes of *both* of the corresponding features.

By plotting the College points sized by their Bayes factor on incomes greater than $110,000, we can see where the colleges lie that have disproportionately high/low proportions of high-income students.

```
# scale factor for coefficients:
f_mult <-
  max(sqrt(tsne_df_all$Y1^2 + tsne_df_all$Y2^2))/
  max(sqrt(tsne_coef_df_all$Y1[-1]^2 + tsne_coef_df_all$Y2[-1]^2))

y2_min <- -3.5
tsne_coef_df_all %>%
  mutate(
    Y2 = pmax(y2_min,Y2*f_mult),
    Y1 = Y1*f_mult,
    mag = sqrt(Y1^2 + Y2^2),
    Coefficient = gsub('\\([^)]+\\)|(_*2005)|_','',gsub('BF_','',Coefficient))
  ) %>%
  {
    ggplot(., aes( x = Y1, y = Y2 ) ) +
      geom_point( color = 'red', alpha = 0.1 ) +
      # Labels for the coefficients
      geom_text(
        aes( label = Coefficient),
        color = 'red',
        alpha = 0.7,
        size = 3,
        check_overlap = TRUE
      ) +
      # Rays on the coefficients
      geom_segment(
        inherit.aes = FALSE,
        data = (.) %>% filter(mag>1),
        aes( x=0, y=0, xend=Y1, yend=Y2 ),
        color = 'red',
        alpha = 0.3,
        arrow = arrow(length = unit(0.03, "npc"))
      ) +
      # Labels for the Colleges
      geom_text(
        inherit.aes = FALSE,
        data = tsne_df_all,
        aes( x=Y1, y=Y2, label=College ),
        mapping=,
        color = 'black',
        size=3,
```

```
      check_overlap = TRUE
    ) +
    # Points for the colleges
    geom_point( data=tsne_df_all, aes(x=Y1,y=Y2, size = BF_Income_gt110K ), color='blue',alpha=0.1) +
    ggtitle( "t-SNE Biplot" , subtitle = "(blue = college, red = feature)") +
    theme( text = element_text( face = 'bold' ) ) #+
    #scale_y_continuous(limits = c(y2_min,4))
  #scale_y_continuous(limits = c(y2_min,4))
} %>%
print()
```

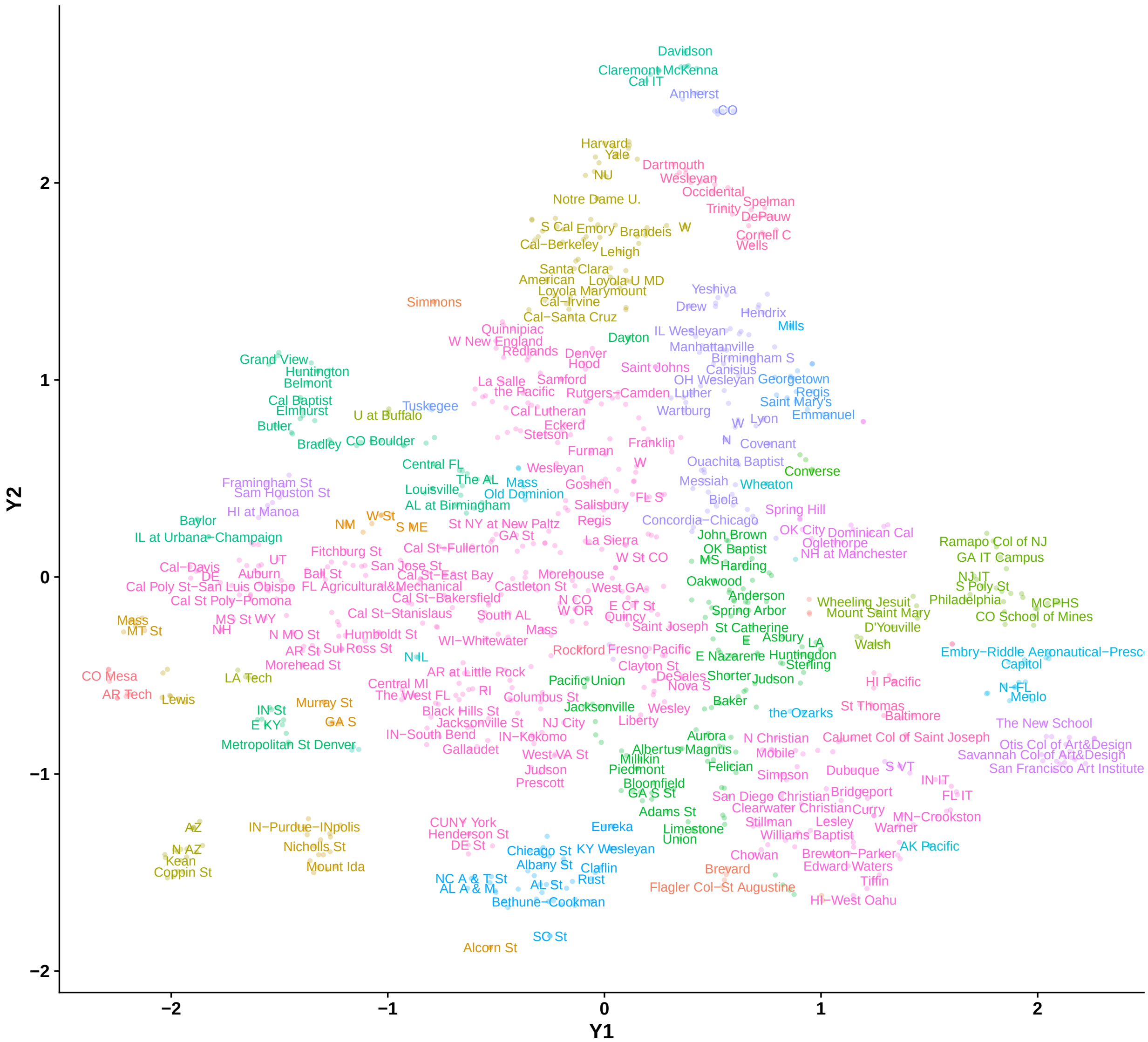


Perform Hierarchical Clustering

Now, I perform cluster analysis. Hierarchical clustering is a quick way to identify clusters in the 2-D t-SNE space. We can then color the clusterings in a scatterplot to more easily visualize the structure.

```r
tsne_mat_hc_all <- tsne_df_all %>% select(Y1,Y2) %>% as.matrix() %>% set_rownames(tsne_df_all$College)
hc_all <- hclust( d = dist( tsne_mat_hc_all ), method = 'single' )
n_cluster <- 55
cluster_id_all <- cutree( hc_all, k = n_cluster )

# plot( tsne_mat_hc, pch=20, cex=0.5 )
# for(j in seq_along(cl)){
#   points( tsne_mat_hc[ cl[[j]], ], pch=20, col=j, cex=1)
# }

# randomize so adjacent clusters are more likely to have very different colors.
set.seed(137)
cluster_id_all <- setNames( sample.int(n_cluster)[cluster_id_all], names(cluster_id_all) )
```

```r
tsne_mat_hc_all %>%
  as_tibble() %>%
  mutate( College = names(cluster_id_all), cluster = factor( cluster_id_all ) ) %>%
  {
    ggplot(.,aes( x = Y1, y = Y2, color = cluster ) ) +
      geom_point( size = 1, alpha = 0.3 ) +
      geom_text( aes(label = College ), size = 3, check_overlap = TRUE ) +
      theme(
        text = element_text( face = 'bold' ),
        legend.position = 'none'
      )
  } %>%
  print()
```



Characterize the Clusters

For each cluster, calculate its mutual information with each (discretized) Bayes factor.

```
fctr_clstr <- factor( sprintf( "C%02d", cluster_id_all ) )
# df_cluster <- DataSpec$studentBF %>%
#   dplyr::select( College,one_of( unlist(strsplit(key_terms,":")) ) %>%
#   mutate( cluster = fctr_clstr ) %>%
#   gather( key= Feature, value = Value, -cluster, -College )

key_features <- unique(unlist(strsplit(tsne_coef_df_all$Coefficient,":")))

df_cluster <- mmat %>%
  as_tibble() %>%
  dplyr::select( one_of( key_features ) ) %>%
  bind_cols(DataSpec$studentBF %>% dplyr::select(unitID,College)) %>%
  mutate(cluster = fctr_clstr) %>%
```

```r
  gather( key= Feature, value = Value, one_of( key_features ) ) %>%
  mutate( Feature = gsub('BF_','',Feature ) )

df_median <- df_cluster %>%
  group_by(Feature,cluster) %>%
  summarize_at(funs(median),.vars=vars(Value)) %>%
  ungroup()

#set.seed(131)
college_label <- DataSpec$studentBF %>%
  mutate(
    cluster = fctr_clstr,
    shortname = college_names,
    Y1 = tsne_df_all$Y1,
    Y2 = tsne_df_all$Y2
  ) %>%
  group_by(cluster) %>%
  summarize(
    i_max = which.max(BF_SAT_gt1400),
    name_max_SAT   = shortname[i_max],
    unitID_max_SAT = unitID[i_max],
    Y1_max = Y1[i_max],
    Y2_max = Y2[i_max]
  )

nm_max <- college_label %$% { setNames( name_max_SAT, as.character(cluster) ) }
Y1_max <- college_label %$% { setNames( Y1_max, as.character(cluster) ) }
Y2_max <- college_label %$% { setNames( Y2_max, as.character(cluster) ) }

df_mi <- df_cluster %>%
  left_join( college_label, by = 'cluster' ) %>%
  group_by( Feature ) %>%
  do(
    {
      feature <- (.)$Feature[[1]]
      sapply(
        levels(fctr_clstr),
        function(clstr, min_lvl, nm_max, Y1_max, Y2_max ) {
          c(
            name_max_SAT = nm_max[[clstr]],
            Y1_max  = Y1_max[[clstr]],
            Y2_max  = Y2_max[[clstr]],
            Cluster = clstr,
            median  = df_median %>% filter(Feature == feature, cluster == clstr ) %$% Value,
            mi      = (.) %$%
              mutinformation(
                cluster == clstr,
                if( length( unique(Value) ) <= min_lvl ) {
                  as.character(Value)
                } else {
                  discretize( data.frame( Value = Value ) )
                }
              )
```

```
          )
        },
        min_lvl = 5,
        nm_max = nm_max, Y1_max = Y1_max, Y2_max = Y2_max
      ) %>%
        t() %>%
        as_data_frame() %>%
        mutate(
          Feature = feature,
          median  = as.double(median),
          mi      = as.double(mi),
          Cluster = factor(Cluster)
        ) %>%
        dplyr::select( Feature, Cluster, median, mi, name_max_SAT )
    }
  ) %>% ungroup()

df_mi_rel <- df_mi %>%
  group_by(Cluster) %>%
  mutate(is_max = mi %>% {. == max(.)}) %>%
  ungroup() %>%
  mutate( mi_rel = round(sign(median)*mi/max(mi),2) )  %>%
  arrange(Cluster, desc(mi) ) %>%
  dplyr::select( Cluster, Feature, mi_rel, everything() )

df_mi %<>%
  left_join( df_mi_rel %>% dplyr::select( 1:3 ), by = c('Cluster','Feature') )%>%
  mutate( Cluster_label = sprintf("%s (%s:{%4.1f,%4.1f})", Cluster, name_max_SAT,Y1_max,Y2_max ) )
```

Print a table of the features that most strongly characterize each cluster.

```
df_mi_rel %>%
  filter(mi>=0.02 | is_max ) %>%
  mutate_at(funs(round(.,2)),.vars=vars(median,mi)) %>%
  print(n=Inf)
```

```
## # A tibble: 239 x 7
##     Cluster                    Feature mi_rel median    mi
##      <fctr>                      <chr>  <dbl>  <dbl> <dbl>
##   1     C01          TheologyReligious   0.05   2.12  0.01
##   2     C02           VisualPerforming   0.06   0.54  0.01
##   3     C03             FamilyConsumer   0.04   1.66  0.00
##   4     C04                discBreadth  -0.04  -0.98  0.00
##   5     C05         BusinessManagement   0.09   0.25  0.01
##   6     C06 CommunicationsTechnologies   0.14   3.55  0.01
##   7     C07            EnglishLanguage   0.04   0.01  0.00
##   8     C07                 gt24yrsold   0.04   0.68  0.00
##   9     C08                Engineering   0.08   1.33  0.01
##  10     C09            PersonalCulinary   0.61   6.42  0.07
##  11     C10             MechanicRepair   0.26   7.48  0.03
##  12     C11          C150_4_POOLED_SUPP   1.00   0.73  0.11
##  13     C11              fsend_2_2005  -0.99  -2.03  0.11
##  14     C11              fsend_5_2005   0.96   1.76  0.10
##  15     C11             SAT_gt800le1000  -0.92  -1.50  0.10
```

```
## 16      C11                   SAT_le800  -0.87  -2.06  0.09
## 17      C11                  SAT_gt1400   0.83   1.18  0.09
## 18      C11                  RPY_7YR_RT   0.80   0.18  0.08
## 19      C11                     CDR3est  -0.79  -1.59  0.08
## 20      C11                p_gt75Kle110K   0.71   1.25  0.08
## 21      C11            SAT_gt1200le1400   0.68   0.88  0.07
## 22      C11              pell_ever_2005  -0.68  -1.61  0.07
## 23      C11                    ADM_RATE  -0.67  -1.43  0.07
## 24      C11                  RPY_5YR_RT   0.64   0.19  0.07
## 25      C11              SocialSciences   0.59   0.54  0.06
## 26      C11                 fsend_1_2005  -0.51  -1.24  0.05
## 27      C11                  gt24yrsold  -0.46  -0.75  0.05
## 28      C11                  AreaEthnic   0.43   1.36  0.05
## 29      C11                   Education  -0.43  -0.05  0.05
## 30      C11                 p_gt48Kle75K   0.39   0.93  0.04
## 31      C11         PhilosophyReligious   0.35   0.75  0.04
## 32      C11            ForeignLanguages   0.34   0.77  0.04
## 33      C11                     veteran   0.26   0.37  0.03
## 34      C11       MathematicsStatistics   0.25   0.49  0.03
## 35      C11                 Engineering   0.25   1.35  0.03
## 36      C11            PhysicalSciences   0.23   0.57  0.02
## 37      C11             ParksRecreation  -0.20  -1.19  0.02
## 38      C11            HomelandSecurity  -0.20  -0.92  0.02
## 39      C12         ScienceTechnologies   0.72   7.30  0.08
## 40      C13     TransportationMaterials   0.10   4.15  0.01
## 41      C14         ArchitectureRelated   0.06   2.40  0.01
## 42      C15            VisualPerforming  -0.23  -2.96  0.02
## 43      C16             EnglishLanguage  -0.47  -3.39  0.05
## 44      C16                 p_gt48Kle75K   0.32   1.70  0.03
## 45      C16                   Education  -0.29  -1.83  0.03
## 46      C16                p_gt75Kle110K   0.29   1.76  0.03
## 47      C16                     History  -0.27  -2.73  0.03
## 48      C16                 discBreadth  -0.27  -1.60  0.03
## 49      C16         PhilosophyReligious  -0.22  -1.43  0.02
## 50      C16              SocialSciences  -0.19  -0.04  0.02
## 51      C17     TransportationMaterials   0.03   4.68  0.00
## 52      C18                     History   0.08   0.58  0.01
## 53      C19              SocialSciences   0.53   0.11  0.06
## 54      C19           TheologyReligious   0.48   1.97  0.05
## 55      C19                  AreaEthnic  -0.41  -0.78  0.04
## 56      C19           C150_4_POOLED_SUPP  -0.35  -0.08  0.04
## 57      C19             SAT_gt800le1000   0.35   0.38  0.04
## 58      C19                 fsend_5_2005  -0.32  -0.51  0.03
## 59      C19            SAT_gt1200le1400  -0.32  -0.07  0.03
## 60      C19                  gt24yrsold   0.32   0.53  0.03
## 61      C19                   SAT_le800   0.31   0.49  0.03
## 62      C19              pell_ever_2005   0.31   0.25  0.03
## 63      C19                     History   0.28   0.28  0.03
## 64      C19                  RPY_7YR_RT   0.26   0.15  0.03
## 65      C19                   Education   0.26   0.60  0.03
## 66      C19           HealthProfessions   0.26   0.64  0.03
## 67      C19                  RPY_5YR_RT   0.25   0.13  0.03
## 68      C19            ForeignLanguages  -0.24  -1.45  0.03
## 69      C19                 fsend_1_2005   0.23   0.57  0.02
```

```
## 70      C19                discBreadth  -0.22  -0.11  0.02
## 71      C19           PhysicalSciences   0.22   0.33  0.02
## 72      C19                SAT_gt1400  -0.21  -0.05  0.02
## 73      C19           NaturalResources  -0.21  -0.90  0.02
## 74      C19                   veteran   0.20   0.48  0.02
## 75      C19                   CDR3est   0.20   0.25  0.02
## 76      C19           ParksRecreation   0.19   0.85  0.02
## 77      C20               p_gt48Kle75K  -0.08  -2.13  0.01
## 78      C21        ComputerInformation   0.04   0.40  0.00
## 79      C22    TransportationMaterials   0.28   4.16  0.03
## 80      C23 CommunicationsTechnologies   0.33   3.09  0.04
## 81      C23         PhilosophyReligious   0.21   0.65  0.02
## 82      C24    TransportationMaterials   0.10   3.73  0.01
## 83      C25           SAT_gt800le1000  -0.24  -3.31  0.03
## 84      C25                  ADM_RATE  -0.24  -3.06  0.03
## 85      C25                gt24yrsold  -0.24  -2.26  0.03
## 86      C25                  SAT_le800  -0.24  -3.82  0.03
## 87      C25          C150_4_POOLED_SUPP   0.23   0.82  0.02
## 88      C25                SAT_gt1400   0.23   1.43  0.02
## 89      C25                   CDR3est  -0.21  -2.46  0.02
## 90      C25                RPY_5YR_RT  -0.21  -7.80  0.02
## 91      C25               fsend_5_2005   0.20   1.99  0.02
## 92      C25               fsend_2_2005  -0.20  -3.31  0.02
## 93      C25                RPY_7YR_RT  -0.19  -6.89  0.02
## 94      C26 CommunicationsTechnologies   0.07   3.57  0.01
## 95      C27               fsend_5_2005  -0.02  -0.22  0.00
## 96      C28                discBreadth   0.04   0.76  0.00
## 97      C28                Engineering   0.04   1.19  0.00
## 98      C28               p_gt30Kle48K   0.04   0.20  0.00
## 99      C29              p_gt75Kle110K   0.04   0.24  0.00
## 100     C30               fsend_1_2005   0.04   1.33  0.00
## 101     C31               p_gt48Kle75K   0.09   0.65  0.01
## 102     C32                discBreadth  -0.38  -2.63  0.04
## 103     C32            EnglishLanguage  -0.35  -3.39  0.04
## 104     C32      MathematicsStatistics  -0.33  -2.50  0.03
## 105     C32                   History  -0.32  -2.73  0.03
## 106     C32            VisualPerforming  -0.31  -2.96  0.03
## 107     C32             SocialSciences  -0.31  -2.77  0.03
## 108     C32                 Education  -0.22  -1.83  0.02
## 109     C32           PhysicalSciences  -0.20  -2.05  0.02
## 110     C33             pell_ever_2005   0.05   0.37  0.00
## 111     C34           ForeignLanguages   0.04   0.77  0.00
## 112     C35             SocialSciences   0.04   0.30  0.00
## 113     C36                RPY_5YR_RT  -0.67  -0.07  0.07
## 114     C36           SAT_gt1200le1400  -0.66  -1.74  0.07
## 115     C36                RPY_7YR_RT   0.62   0.03  0.07
## 116     C36                SAT_gt1400  -0.62  -1.91  0.07
## 117     C36               p_gt48Kle75K  -0.57  -1.85  0.06
## 118     C36             pell_ever_2005   0.54   1.49  0.06
## 119     C36                   CDR3est   0.53   1.58  0.06
## 120     C36                  SAT_le800   0.52   0.83  0.05
## 121     C36              p_gt75Kle110K  -0.42  -1.44  0.04
## 122     C36          C150_4_POOLED_SUPP  -0.40  -0.66  0.04
## 123     C36               p_gt30Kle48K  -0.34  -1.43  0.04
```

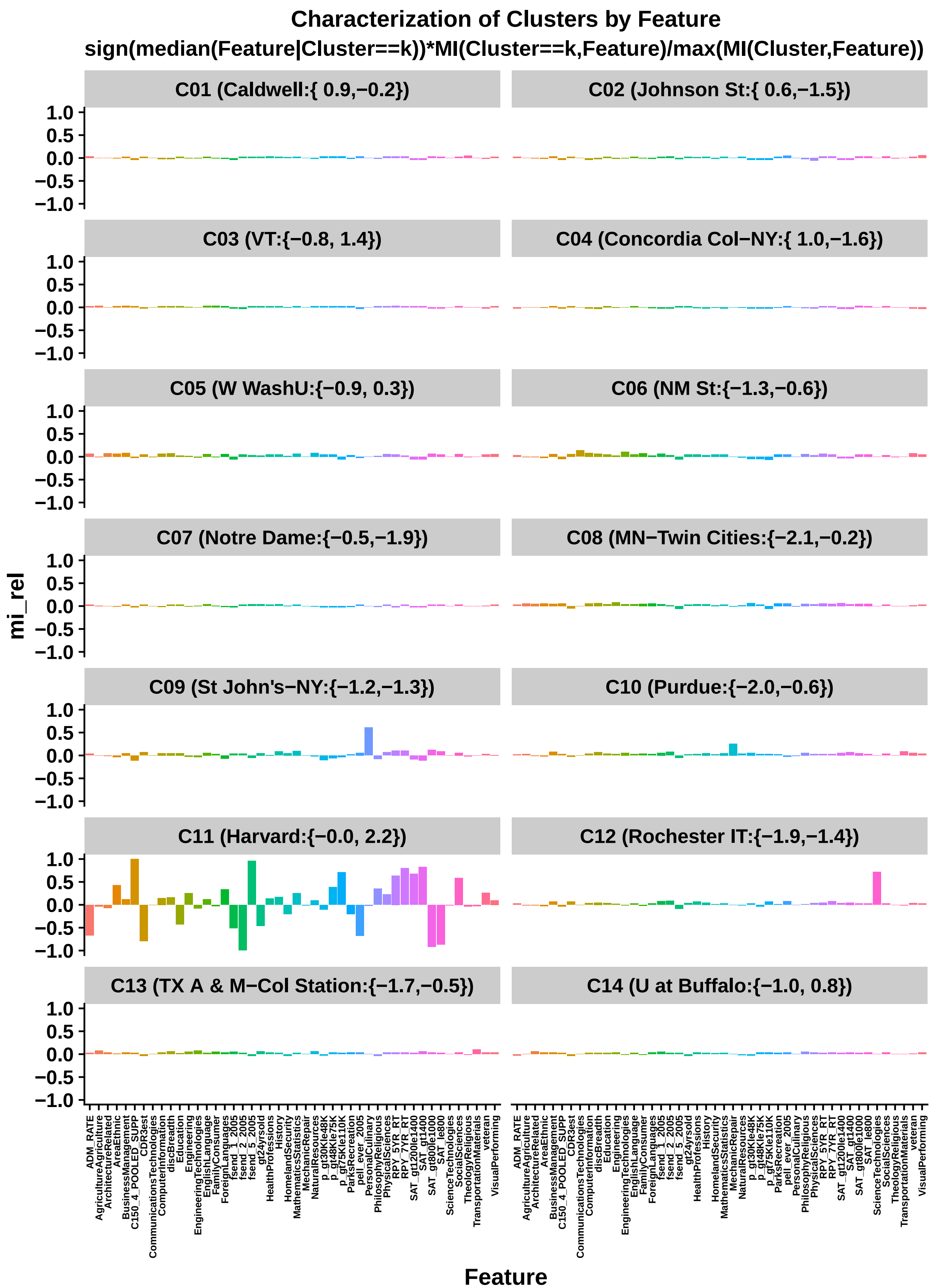
```
## 124      C36            HomelandSecurity   0.26   1.17  0.03
## 125      C36            ForeignLanguages  -0.24  -1.45  0.03
## 126      C36                fsend_1_2005  -0.19  -0.60  0.02
## 127      C37                     veteran  -0.22  -2.10  0.02
## 128      C38       MathematicsStatistics   0.04   0.43  0.00
## 129      C39          BusinessManagement  -0.20  -3.46  0.02
## 130      C39                  gt24yrsold  -0.20  -2.17  0.02
## 131      C39                  RPY_7YR_RT  -0.20  -6.89  0.02
## 132      C39          C150_4_POOLED_SUPP   0.19   0.78  0.02
## 133      C39                  SAT_gt1400   0.19   1.20  0.02
## 134      C39              SocialSciences   0.19   0.62  0.02
## 135      C39                fsend_2_2005  -0.19  -3.62  0.02
## 136      C40                     veteran  -0.51  -2.10  0.05
## 137      C40             SAT_gt800le1000   0.44   0.10  0.05
## 138      C40                  gt24yrsold  -0.41  -0.98  0.04
## 139      C40          C150_4_POOLED_SUPP   0.40   0.37  0.04
## 140      C40            SAT_gt1200le1400   0.39   0.52  0.04
## 141      C40                  RPY_5YR_RT   0.38   0.17  0.04
## 142      C40                  RPY_7YR_RT   0.37   0.17  0.04
## 143      C40              pell_ever_2005  -0.37  -0.83  0.04
## 144      C40                   SAT_le800   0.35   0.00  0.04
## 145      C40            PhysicalSciences   0.34   0.58  0.04
## 146      C40                  SAT_gt1400   0.29   0.59  0.03
## 147      C40                     CDR3est  -0.28  -0.66  0.03
## 148      C40                     History   0.27   0.47  0.03
## 149      C40          PhilosophyReligious   0.26   0.78  0.03
## 150      C40            HomelandSecurity  -0.26  -0.92  0.03
## 151      C40             EnglishLanguage   0.25   0.42  0.03
## 152      C40            ForeignLanguages   0.22   0.73  0.02
## 153      C41                fsend_2_2005   0.06   0.73  0.01
## 154      C42 CommunicationsTechnologies   0.21   3.66  0.02
## 155      C43                 discBreadth  -0.40  -3.92  0.04
## 156      C43                     History  -0.34  -2.73  0.04
## 157      C43            VisualPerforming   0.32   0.90  0.03
## 158      C43              SocialSciences  -0.31  -2.77  0.03
## 159      C43       MathematicsStatistics  -0.31  -2.50  0.03
## 160      C43          BusinessManagement  -0.31  -3.46  0.03
## 161      C43            PhysicalSciences  -0.29  -2.05  0.03
## 162      C43         ComputerInformation  -0.28  -1.97  0.03
## 163      C43             EnglishLanguage  -0.26  -3.39  0.03
## 164      C43            ForeignLanguages  -0.20  -1.45  0.02
## 165      C44            VisualPerforming  -0.10  -2.96  0.01
## 166      C45                fsend_5_2005   0.18   0.91  0.02
## 167      C46            PhysicalSciences   0.05   0.31  0.01
## 168      C47                  AreaEthnic   0.04   1.35  0.00
## 169      C48            PhysicalSciences  -0.98  -2.05  0.10
## 170      C48            ForeignLanguages  -0.59  -1.45  0.06
## 171      C48                 discBreadth  -0.50  -0.56  0.05
## 172      C48              SocialSciences  -0.43  -0.11  0.05
## 173      C48       MathematicsStatistics  -0.40  -2.50  0.04
## 174      C48          C150_4_POOLED_SUPP  -0.32  -0.29  0.03
## 175      C48                   SAT_le800   0.29   0.56  0.03
## 176      C48         ComputerInformation  -0.29  -1.97  0.03
## 177      C48             SAT_gt800le1000   0.25   0.40  0.03
```

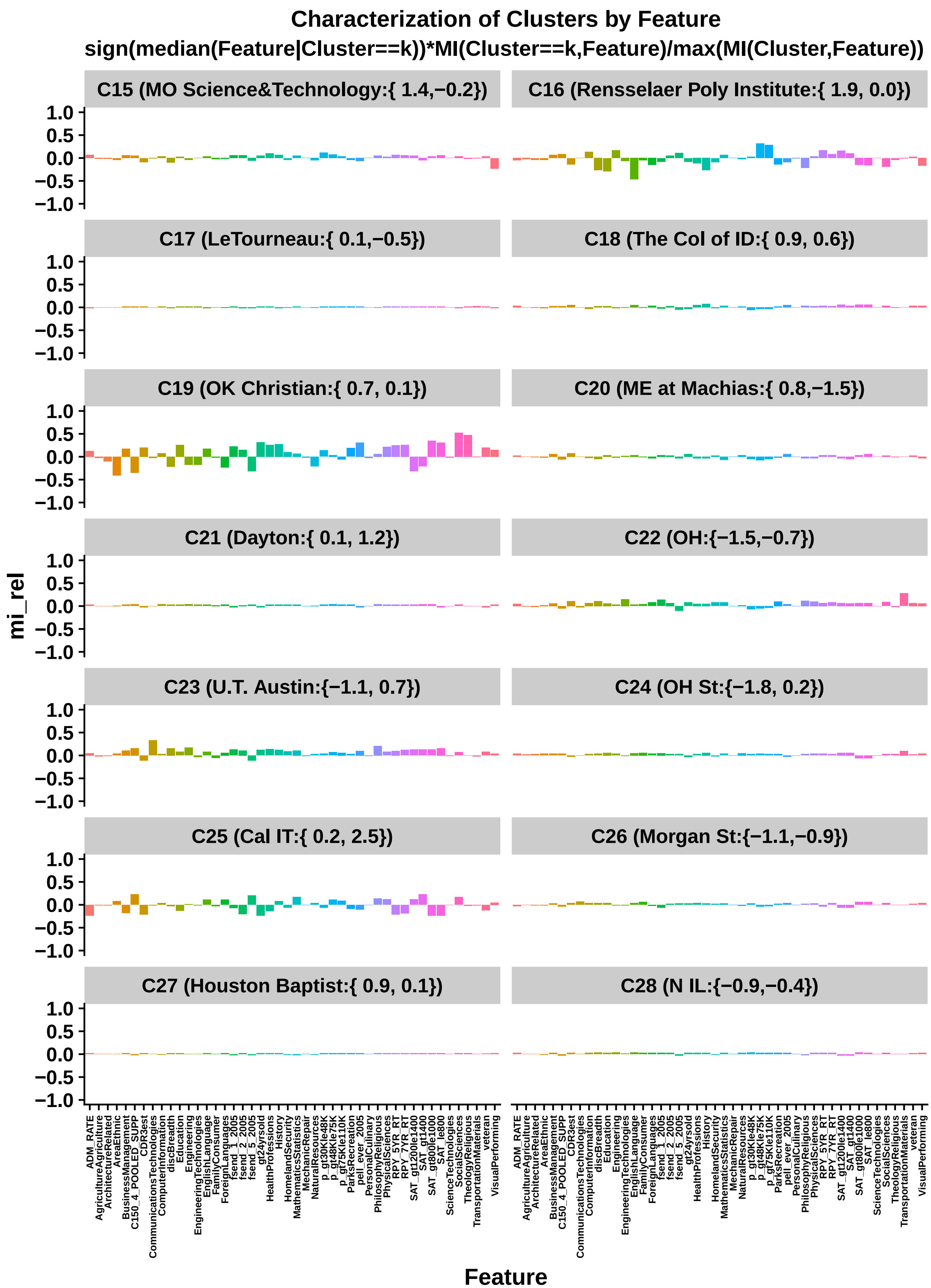
```
## 178      C48            SAT_gt1200le1400  -0.23  -0.27  0.02
## 179      C48                     History   0.23   0.16  0.02
## 180      C48                  RPY_5YR_RT   0.23   0.11  0.02
## 181      C48                  AreaEthnic  -0.22  -0.78  0.02
## 182      C48          PhilosophyReligious  -0.21  -1.43  0.02
## 183      C48              EnglishLanguage   0.20   0.11  0.02
## 184      C49               SocialSciences   0.67   0.37  0.07
## 185      C49             ForeignLanguages   0.61   0.65  0.06
## 186      C49                 discBreadth   0.57   0.47  0.06
## 187      C49             PhysicalSciences   0.52   0.46  0.06
## 188      C49                     veteran   0.47   0.48  0.05
## 189      C49                     History   0.47   0.36  0.05
## 190      C49        MathematicsStatistics   0.46   0.37  0.05
## 191      C49          PhilosophyReligious   0.42   0.61  0.04
## 192      C49                  SAT_gt1400   0.41   0.04  0.04
## 193      C49              EnglishLanguage   0.38   0.31  0.04
## 194      C49           BusinessManagement   0.37   0.30  0.04
## 195      C49             VisualPerforming   0.34   0.34  0.04
## 196      C49                 fsend_2_2005   0.33   0.33  0.04
## 197      C49              SAT_gt800le1000   0.29   0.33  0.03
## 198      C49                 fsend_5_2005  -0.29  -0.36  0.03
## 199      C49            SAT_gt1200le1400   0.29   0.14  0.03
## 200      C49            HealthProfessions   0.26   0.57  0.03
## 201      C49                  gt24yrsold   0.24   0.33  0.03
## 202      C49          ComputerInformation   0.24   0.49  0.03
## 203      C49                  RPY_5YR_RT   0.23   0.14  0.02
## 204      C49              ParksRecreation   0.23   0.82  0.02
## 205      C49                    SAT_le800   0.22   0.35  0.02
## 206      C49 CommunicationsTechnologies  -0.19  -0.27  0.02
## 207      C50     TransportationMaterials   0.15   4.25  0.02
## 208      C51             VisualPerforming  -0.20  -2.96  0.02
## 209      C52           BusinessManagement  -0.57  -3.46  0.06
## 210      C52               SocialSciences   0.43   0.61  0.05
## 211      C52                  gt24yrsold  -0.36  -1.63  0.04
## 212      C52              EnglishLanguage   0.35   0.51  0.04
## 213      C52             PhysicalSciences   0.35   0.67  0.04
## 214      C52                 fsend_5_2005   0.34   1.74  0.04
## 215      C52          PhilosophyReligious   0.31   0.84  0.03
## 216      C52                 fsend_2_2005  -0.31  -1.87  0.03
## 217      C52                  RPY_7YR_RT   0.29   0.18  0.03
## 218      C52            SAT_gt1200le1400   0.28   0.91  0.03
## 219      C52           C150_4_POOLED_SUPP   0.28   0.64  0.03
## 220      C52              SAT_gt800le1000  -0.28  -0.83  0.03
## 221      C52                  AreaEthnic   0.27   1.47  0.03
## 222      C52            HealthProfessions  -0.27  -1.72  0.03
## 223      C52                     veteran  -0.26  -2.10  0.03
## 224      C52                    SAT_le800  -0.25  -1.25  0.03
## 225      C52             ForeignLanguages   0.25   0.83  0.03
## 226      C52                  SAT_gt1400   0.25   0.95  0.03
## 227      C52                  RPY_5YR_RT   0.24   0.18  0.03
## 228      C52                     History   0.23   0.50  0.02
## 229      C52                 fsend_1_2005  -0.22  -1.13  0.02
## 230      C52        MathematicsStatistics   0.22   0.56  0.02
## 231      C52                     CDR3est  -0.21  -1.20  0.02
```

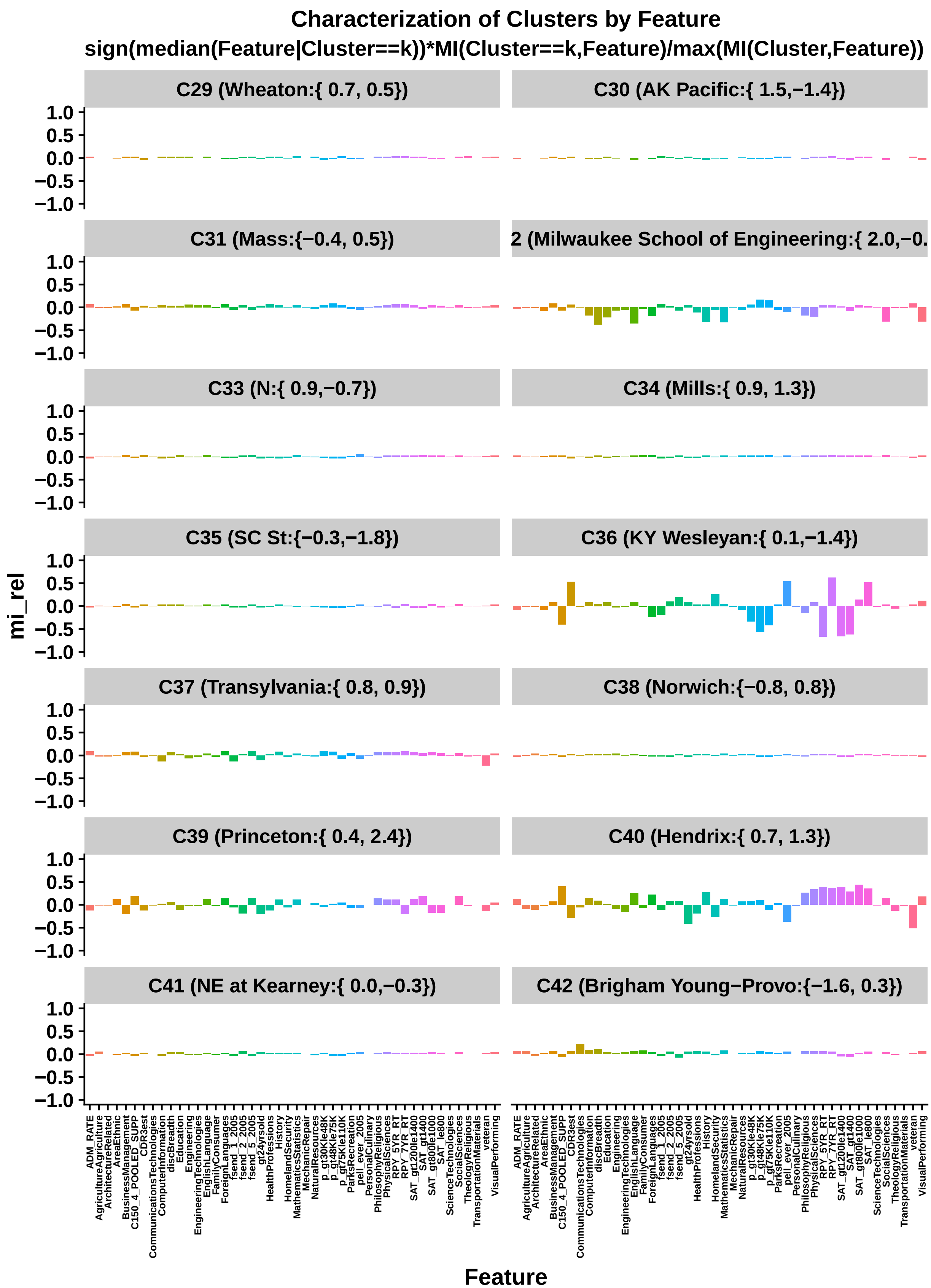
```
## 232     C52             pell_ever_2005  -0.21  -1.06  0.02
## 233     C53          C150_4_POOLED_SUPP  -0.04  -0.15  0.00
## 234     C53               p_gt30Kle48K  -0.04  -0.10  0.00
## 235     C53               p_gt48Kle75K   0.04   0.21  0.00
## 236     C53              p_gt75Kle110K   0.04   0.45  0.00
## 237     C53                   SAT_le800  -0.04  -0.52  0.00
## 238     C54         ComputerInformation   0.04   0.49  0.00
## 239     C55              MechanicRepair   0.34   8.27  0.04
## # ... with 2 more variables: name_max_SAT <chr>, is_max <lgl>
```

Plot all features of each cluster.

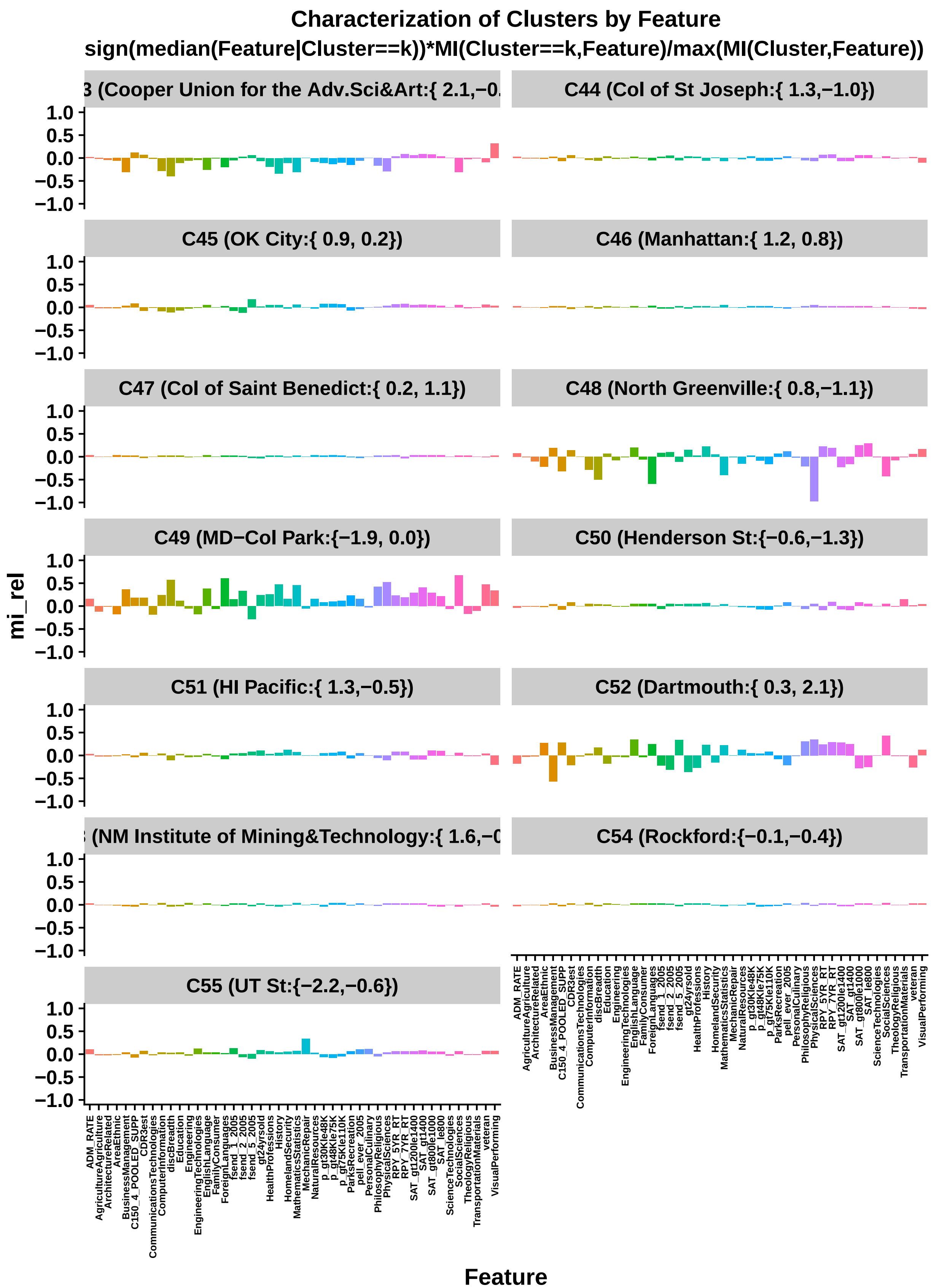
```
i_clstr_min <- 0L
n_per_plot <- 14L
n_seq <- fctr_clstr %>% nlevels() %>% {seq(n_per_plot,.,length.out = ./n_per_plot) %>% ceiling()} %>% fl
for( i_clstr_max in n_seq ){
  df_mi %>% filter(  as.integer(Cluster) > i_clstr_min, as.integer(Cluster) <= i_clstr_max ) %>%
  {
    ggplot(., aes( x = Feature, y = mi_rel, fill = Feature ) ) +
      geom_bar( stat = 'identity', position = 'dodge' ) +
      ylim( c(-1,1) ) +
      facet_wrap( ~ Cluster_label, nrow = 7, ncol = 2 ) +
      ggtitle(
        label = "Characterization of Clusters by Feature",
        subtitle = "sign(median(Feature|Cluster==k))*MI(Cluster==k,Feature)/max(MI(Cluster,Feature))") +
      theme(
        text = element_text( face = 'bold' ),
        axis.text.x = element_text(angle=90,hjust=1,vjust=0.5, size = 6 ),
        legend.position = 'none'
      )
  } %>%
    print()
  i_clstr_min <- i_clstr_max
}
```


Characterization of Clusters by Feature
sign(median(Feature|Cluster==k))*MI(Cluster==k,Feature)/max(MI(Cluster,Feature))
C01 (Caldwell:{ 0.9,−0.2})
C02 (Johnson St:{ 0.6,−1.5})
C03 (VT:{−0.8, 1.4})
C04 (Concordia Col−NY:{ 1.0,−1.6})
C05 (W WashU:{−0.9, 0.3})
C06 (NM St:{−1.3,−0.6})
C07 (Notre Dame:{−0.5,−1.9})
C08 (MN−Twin Cities:{−2.1,−0.2})
C09 (St John's−NY:{−1.2,−1.3})
C10 (Purdue:{−2.0,−0.6})
C11 (Harvard:{−0.0, 2.2})
C12 (Rochester IT:{−1.9,−1.4})
C13 (TX A & M−Col Station:{−1.7,−0.5})
C14 (U at Buffalo:{−1.0, 0.8})
mi_rel
1.0
0.5
0.0
−0.5
−1.0
ADM_RATE
AgricultureAgriculture
ArchitectureRelated
AreaEthnic
BusinessManagement
C150_4_POOLED_SUPP
CDR3est
CommunicationsTechnologies
ComputerInformation
discBreadth
Education
Engineering
EngineeringTechnologies
EnglishLanguage
FamilyConsumer
ForeignLanguages
fsend_1_2005
fsend_2_2005
fsend_5_2005
gt24yrsold
HealthProfessions
History
HomelandSecurity
MathematicsStatistics
MechanicRepair
NaturalResources
p_gt30Kle48K
p_gt48Kle75K
p_gt75Kle110K
ParksRecreation
pell_ever_2005
PersonalCulinary
PhilosophyReligious
PhysicalSciences
RPY_5YR_RT
RPY_7YR_RT
SAT_gt1200le1400
SAT_gt1400
SAT_gt800le1000
SAT_le800
ScienceTechnologies
SocialSciences
TheologyReligious
TransportationMaterials
veteran
VisualPerforming
Feature


Characterization of Clusters by Feature
sign(median(Feature|Cluster==k))*MI(Cluster==k,Feature)/max(MI(Cluster,Feature))
C15 (MO Science&Technology:{ 1.4,−0.2})
C16 (Rensselaer Poly Institute:{ 1.9, 0.0})
C17 (LeTourneau:{ 0.1,−0.5})
C18 (The Col of ID:{ 0.9, 0.6})
C19 (OK Christian:{ 0.7, 0.1})
C20 (ME at Machias:{ 0.8,−1.5})
C21 (Dayton:{ 0.1, 1.2})
C22 (OH:{−1.5,−0.7})
C23 (U.T. Austin:{−1.1, 0.7})
C24 (OH St:{−1.8, 0.2})
C25 (Cal IT:{ 0.2, 2.5})
C26 (Morgan St:{−1.1,−0.9})
C27 (Houston Baptist:{ 0.9, 0.1})
C28 (N IL:{−0.9,−0.4})
mi_rel
1.0
0.5
0.0
−0.5
−1.0
ADM_RATE
AgricultureAgriculture
ArchitectureRelated
AreaEthnic
BusinessManagement
C150_4_POOLED_SUPP
CDR3est
CommunicationsTechnologies
ComputerInformation
discBreadth
Education
Engineering
EngineeringTechnologies
EnglishLanguage
FamilyConsumer
ForeignLanguages
fsend_1_2005
fsend_2_2005
fsend_5_2005
gt24yrsold
HealthProfessions
History
HomelandSecurity
MathematicsStatistics
MechanicRepair
NaturalResources
p_gt30Kle48K
p_gt48Kle75K
p_gt75Kle110K
ParksRecreation
pell_ever_2005
PersonalCulinary
PhilosophyReligious
PhysicalSciences
RPY_5YR_RT
RPY_7YR_RT
SAT_gt1200le1400
SAT_gt1400
SAT_gt800le1000
SAT_le800
ScienceTechnologies
SocialSciences
TheologyReligious
TransportationMaterials
veteran
VisualPerforming
Feature



Characterization of Clusters by Feature
sign(median(Feature|Cluster==k))*MI(Cluster==k,Feature)/max(MI(Cluster,Feature))
C29 (Wheaton:{ 0.7, 0.5})
C30 (AK Pacific:{ 1.5,−1.4})
C31 (Mass:{−0.4, 0.5})
2 (Milwaukee School of Engineering:{ 2.0,−0.
C33 (N:{ 0.9,−0.7})
C34 (Mills:{ 0.9, 1.3})
C35 (SC St:{−0.3,−1.8})
C36 (KY Wesleyan:{ 0.1,−1.4})
C37 (Transylvania:{ 0.8, 0.9})
C38 (Norwich:{−0.8, 0.8})
C39 (Princeton:{ 0.4, 2.4})
C40 (Hendrix:{ 0.7, 1.3})
C41 (NE at Kearney:{ 0.0,−0.3})
C42 (Brigham Young−Provo:{−1.6, 0.3})
1.0
0.5
0.0
−0.5
−1.0
mi_rel
ADM_RATE
AgricultureAgriculture
ArchitectureRelated
AreaEthnic
BusinessManagement
C150_4_POOLED_SUPP
CDR3est
CommunicationsTechnologies
ComputerInformation
discBreadth
Education
Engineering
EngineeringTechnologies
EnglishLanguage
FamilyConsumer
ForeignLanguages
fsend_1_2005
fsend_2_2005
fsend_5_2005
gt24yrsold
HealthProfessions
History
HomelandSecurity
MathematicsStatistics
MechanicRepair
NaturalResources
p_gt30Kle48K
p_gt48Kle75K
p_gt75Kle110K
ParksRecreation
pell_ever_2005
PersonalCulinary
PhilosophyReligious
PhysicalSciences
RPY_5YR_RT
RPY_7YR_RT
SAT_gt1200le1400
SAT_gt1400
SAT_gt800le1000
SAT_le800
ScienceTechnologies
SocialSciences
TheologyReligious
TransportationMaterials
veteran
VisualPerforming
Feature


Characterization of Clusters by Feature
sign(median(Feature|Cluster==k))*MI(Cluster==k,Feature)/max(MI(Cluster,Feature))
3 (Cooper Union for the Adv.Sci&Art:{ 2.1,−0.
C44 (Col of St Joseph:{ 1.3,−1.0})
C45 (OK City:{ 0.9, 0.2})
C46 (Manhattan:{ 1.2, 0.8})
C47 (Col of Saint Benedict:{ 0.2, 1.1})
C48 (North Greenville:{ 0.8,−1.1})
C49 (MD−Col Park:{−1.9, 0.0})
C50 (Henderson St:{−0.6,−1.3})
C51 (HI Pacific:{ 1.3,−0.5})
C52 (Dartmouth:{ 0.3, 2.1})
(NM Institute of Mining&Technology:{ 1.6,−(
C54 (Rockford:{−0.1,−0.4})
C55 (UT St:{−2.2,−0.6})
mi_rel
1.0
0.5
0.0
−0.5
−1.0
ADM_RATE
AgricultureAgriculture
ArchitectureRelated
AreaEthnic
BusinessManagement
C150_4_POOLED_SUPP
CDR3est
CommunicationsTechnologies
ComputerInformation
discBreadth
Education
Engineering
EngineeringTechnologies
EnglishLanguage
FamilyConsumer
ForeignLanguages
fsend_1_2005
fsend_2_2005
fsend_5_2005
gt24yrsold
HealthProfessions
History
HomelandSecurity
MathematicsStatistics
MechanicRepair
NaturalResources
p_gt30Kle48K
p_gt48Kle75K
p_gt75Kle110K
ParksRecreation
pell_ever_2005
PersonalCulinary
PhilosophyReligious
PhysicalSciences
RPY_5YR_RT
RPY_7YR_RT
SAT_gt1200le1400
SAT_gt1400
SAT_gt800le1000
SAT_le800
ScienceTechnologies
SocialSciences
TheologyReligious
TransportationMaterials
veteran
VisualPerforming
Feature

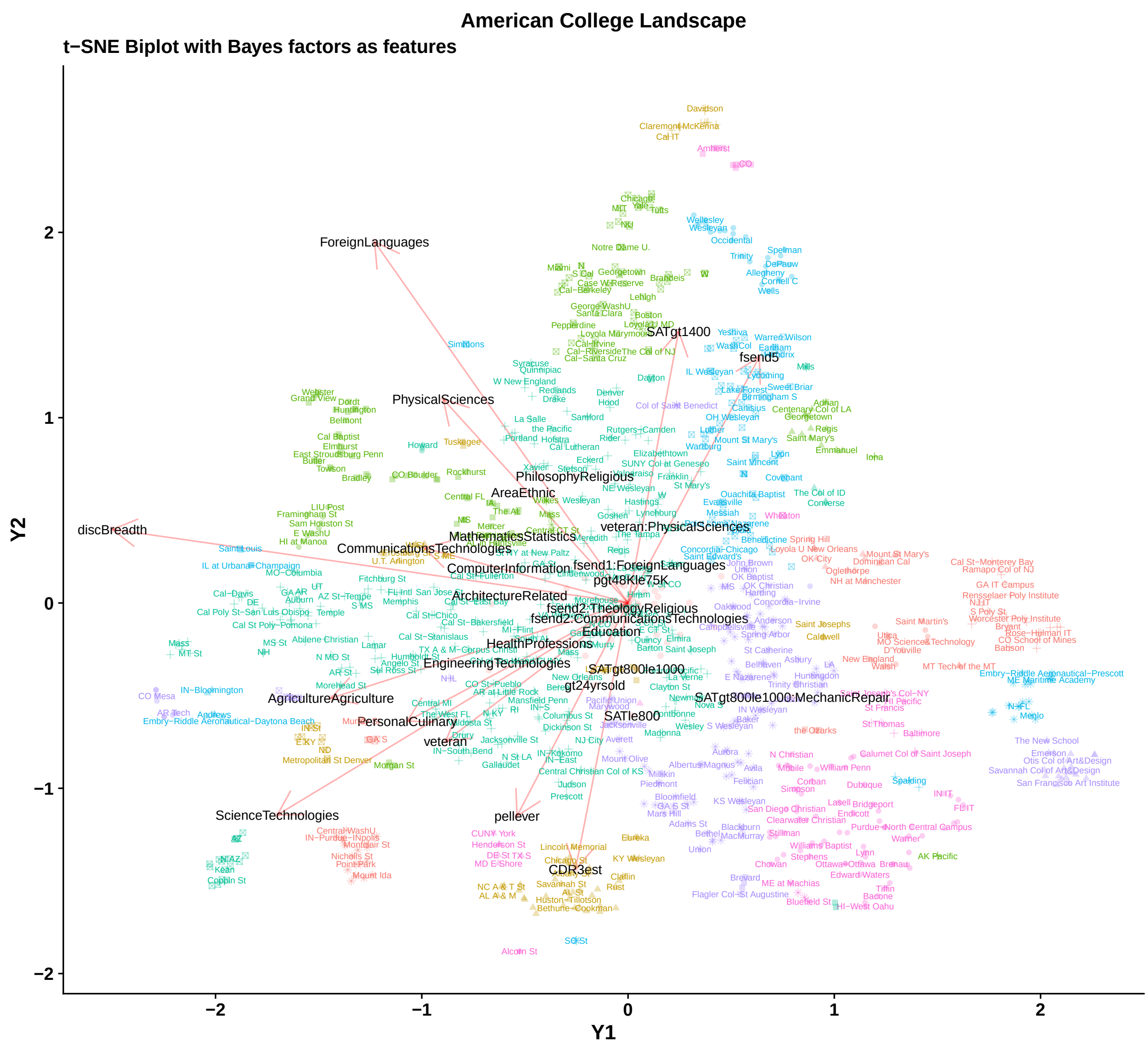
Show Biplot with Cluster Coloring

Finally, we can overlay the feature dimensions on the 2-D plot with cluster coloring.

```r
# Get cluster id for `n_cluster` number of clusters.
cluster_id_all <- cutree( hc_all, k = n_cluster )

# Determine bounds of coordinates for plot.
y2_min <- -4
y2_max <- 3.49
y1 <- range(tsne_mat_hc_all[,1])
y1[1] <- 0.5*floor(y1[1]/0.5)
y1[2] <- 0.5*ceiling(y1[2]/0.5)
y2 <- range(tsne_mat_hc_all[,2])
y2[1] <- 0.5*floor(y2[1]/0.5)
y2[2] <- 0.5*ceiling(y2[2]/0.5)

is_out_of_bounds <- function(x,bounds){ x<bounds[1] | x>bounds[2] }
# Assumes that value violating bounds is of same sign as bound violated AND that bounds are of opposite
bound_factor <- function(x,bounds){
  f1 <- ifelse(x<bounds[1],x/bounds[1],0)
  f2 <- ifelse(x>bounds[2],x/bounds[2],0)
  mapply(function(b1,b2) if(b1>b2) c(1,b1) else c(2,b2),f1,f2)
}
tsne_modified <- tsne_coef_df_all %>%
  mutate(
    Coefficient = gsub('\\([^)]+\\)|(_*2005)|_','',gsub('BF_','',Coefficient)),
    Y1  = f_mult*Y1,
    Y2  = f_mult*Y2 ,
    mag = sqrt(Y1^2 + Y2^2)
  )

# check bounds to find if any violated
bchk1 <- bound_factor(tsne_modified$Y1,y1)
bchk2 <- bound_factor(tsne_modified$Y2,y2)
# bound on Y1 violated
w1 <- which(bchk1[2,] != 0)
# bound on Y2 violated
w2 <- which(bchk2[2,] != 0)
# Keep only coord Y1 or Y2 violated the most by each violating pt.
for( i in intersect(w1,w2)) { if(bchk1[2,i]>bchk2[2,i]) w2<-setdiff(w2,i) else w1<- setdiff(w1,i) }
# bound on Y1 violated: fix it
for( i in w1 ){
  tsne_modified$Y2[i] <- tsne_modified$Y2[i]*y1[bchk1[1,i]]/tsne_modified$Y1[i]
  tsne_modified$Y1[i] <- y1[bchk1[1,i]]
}
# bound on Y2 violated: fix it
for( i in w2 ){
  tsne_modified$Y1[i] <- tsne_modified$Y1[i]*y2[bchk2[1,i]]/tsne_modified$Y2[i]
  tsne_modified$Y2[i] <- y2[bchk2[1,i]]
}
```

```r
# Plot cluster-colored biplot.
tsne_modified %>%
  {
```

```
  ggplot(., aes( x = Y1, y = Y2 ) ) +
    geom_point( color = 'red', alpha = 0.1 ) +
    geom_segment(
      inherit.aes = FALSE,
      data = (.) %>% filter(mag>1),
      aes( x=0, y=0, xend=Y1, yend=Y2 ),
      color = 'red',
      alpha = 0.3,
      arrow = arrow(length = unit(0.03, "npc"))
    ) +
    geom_text(
      inherit.aes = FALSE,
      data = tsne_mat_hc_all %>%
        as_tibble() %>%
        mutate(
          College = names(cluster_id_all),
          cluster = factor( (cluster_id_all %% 7) + 1 )
        ),
      aes( x=Y1, y=Y2, label=College, color = cluster ),
      mapping=,
      show.legend = FALSE,
      size=2,
      check_overlap = TRUE
    ) +
    geom_text(
      aes( label = Coefficient ),
      color = 'black',
      size = 3,
      check_overlap = TRUE
    )  +
    geom_point(
      data = tsne_mat_hc_all %>%
        as_tibble() %>%
        mutate(
          College = names(cluster_id_all),
          cluster = factor( (cluster_id_all %% 7) + 1 ),
          cluster_shape = factor( (cluster_id_all %% 6) + 1 )
        ),
      aes(x=Y1,y=Y2, color = cluster, shape = cluster_shape ),
      show.legend = FALSE,
      alpha=0.3
    ) +
    ggtitle(
      label    = "American College Landscape",
      subtitle = "t-SNE Biplot with Bayes factors as features"
      ) +
    theme( text = element_text( face = 'bold' ) ) #+
    #scale_y_continuous(limits = c(y2_min,5))
} %>%
print()
```



Graph Alignment: Linear Assignment Problem

The t-SNE coordinates can be mapped to a regular 2-D grid by solving the Linear Assignment Problem [4],[5],[6].

Demonstrate LAP Graph Alignment

We first apply LAP Graph Alignment to a simple problem.

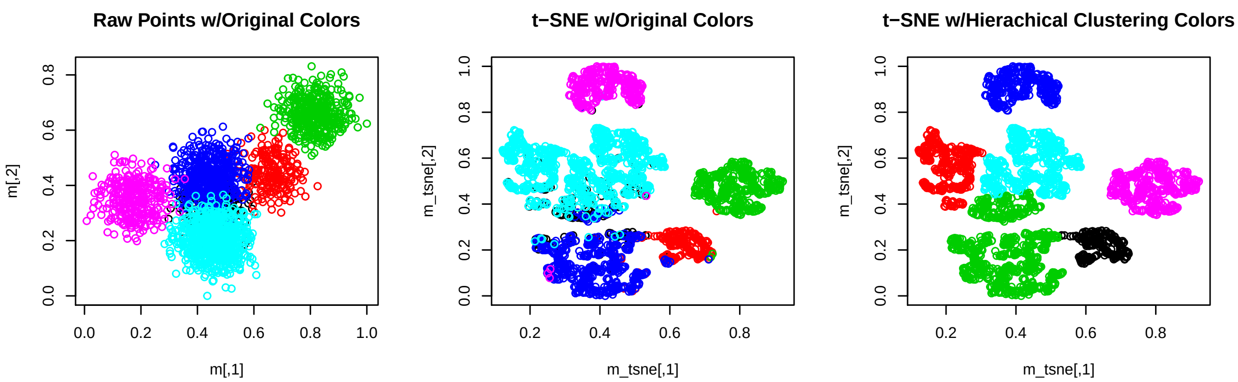
```
set.seed( 13115 )
N_obs <- 50^2
N_clstr <- 6L
mu  <- matrix( rt(N_clstr*2, df = 3),ncol=2 )
```

4 Blog post by by Vadim Markovtsev, 14 March 2017: Jonker-Volgenant Algorithm + t-SNE = Super Powers

5 R. Jonker and A. Volgenant, **"A Shortest Augmenting Path Algorithm for Dense and Sparse Linear Assignment Problems,"** *Computing*, vol. **38**, pp. 325-340, 1987.

6 See: Linear Assignment Problem solver using Jonker-Volgenant algorithm.

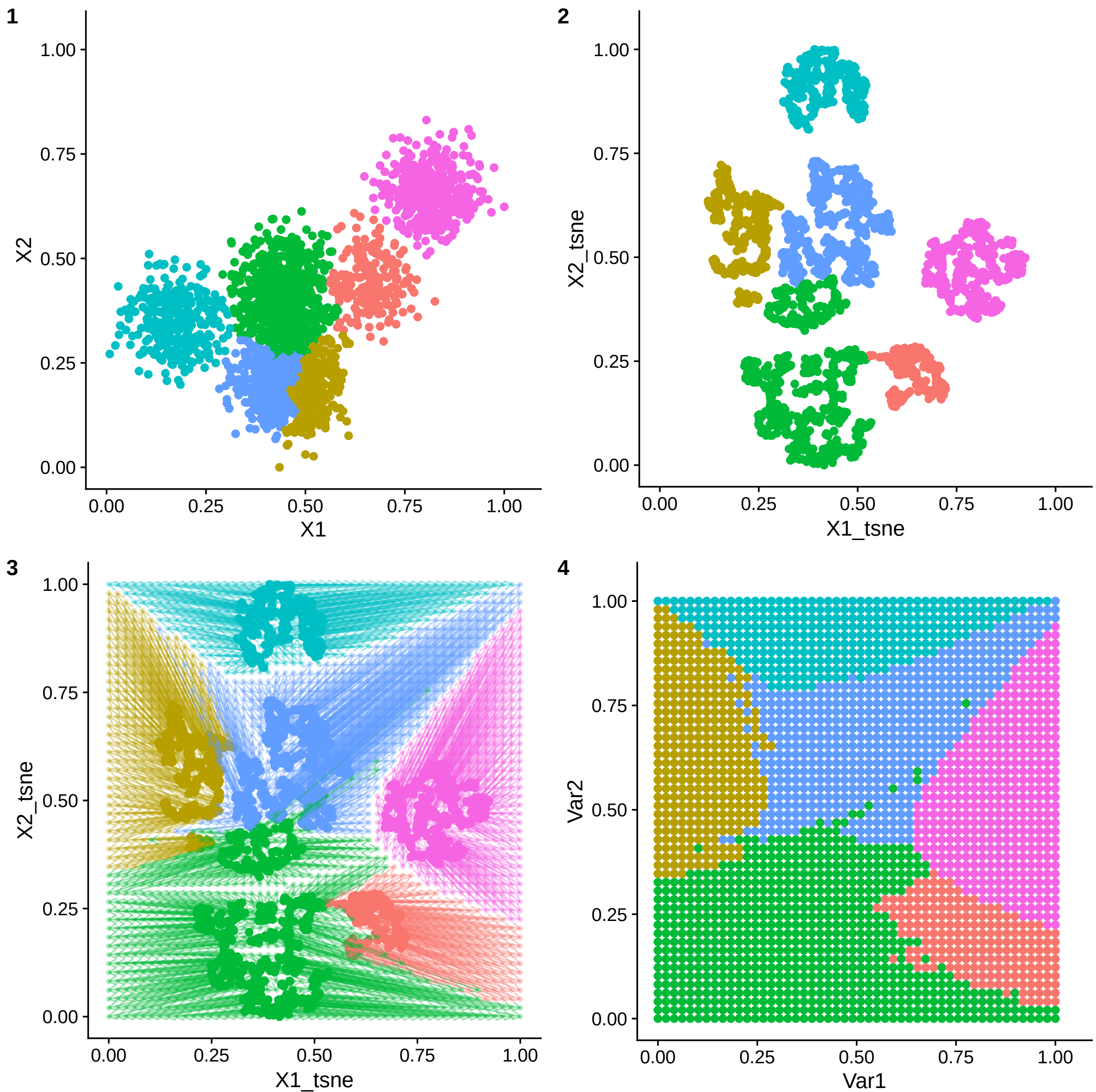
```
p   <- rgamma(N_clstr,3,2) %>% {(.)/sum(.)}
n   <- (p*N_obs) %>% ceiling() %>% {c(N_obs-sum(.[-1]),(.)[-1])}

m   <- n %>%
  seq_along() %>%
  lapply(function(ix) (matrix( rnorm(n[ix]*2,sd=0.5), n[ix], 2 )+matrix(mu[ix,],n[ix],2,byrow=2))) %>%
  {do.call(rbind,.)} %>%
  {((.)-min(.))/diff(range(.))}

par_old <- par( no.readonly = TRUE )
par(mfrow=c(1,3))
plot(m,col=rep(seq_along(n),times=n),main = 'Raw Points w/Original Colors' )

m_tsne <- m %>%
  Rtsne() %$%
  Y %>%
  {((.)-min(.))/diff(range(.))}
plot(m_tsne,col=rep(seq_along(n),times=n), main = 't-SNE w/Original Colors' )

hc  <- m_tsne %>% dist() %>% hclust() %>% cutree(k=N_clstr)
grid <- expand.grid(1:sqrt(N_obs),1:sqrt(N_obs)) %>% as.matrix() %>% {((.)-min(.))/diff(range(.))}
plot(m_tsne,col=hc, main = 't-SNE w/Hierachical Clustering Colors')
```




```
par( par_old )
```

```
cost_matrix <- matrix(NA,nrow(m_tsne),nrow(grid))
for( i in seq_len(nrow(m_tsne))){
  for(j in seq_len(nrow(grid))){
    cost_matrix[i,j] <- sqrt( sum((m_tsne[i,] - grid[j,])^2) )
  }
}
cost_matrix = cost_matrix * (100000 / max(cost_matrix) )

px <- LinearAssignment( cost_matrix )

m_df <- m[px,] %>%
  set_colnames(c("X1","X2") ) %>%
  cbind( m_tsne[px,] %>% set_colnames(c("X1_tsne","X2_tsne") )) %>%
  cbind(grid) %>%
  as_tibble() %>%
```

```r
  mutate( cluster = factor( hc[px] ), row_id = as.character(1:nrow(.)) )

plts <- list(
  original = m_df %>%
  {
    ggplot(.,aes(x=X1,y=X2,color=cluster)) +
      geom_point(size=2) +
      lims( x=c(0,1.04), y=c(0,1.04) ) +
      #geom_text( aes(label = row_id),nudge_y=0.02,size=2) +
      theme( legend.position = 'none' )
  },
  tsne = m_df %>%
  {
    ggplot(.,aes(x=X1_tsne,y=X2_tsne,color=cluster)) +
      geom_point(size=2) +
      lims( x=c(0,1.04), y=c(0,1.04) ) +
      #geom_text( aes(label = row_id),nudge_y=0.02,size=2) +
      theme( legend.position = 'none' )
  },
  assigned = m_df %>%
  {
    ggplot(.,aes(x=X1_tsne,y=X2_tsne,color=cluster) ) +
      geom_point(size=2 ) +
      geom_segment(
        aes(xend=Var1,yend=Var2),
        arrow = arrow( length = unit(0.2,"cm") ),
        #color = 'gray',
        alpha = 0.4
      ) +
      geom_point( aes(x=Var1,y=Var2), size=1, alpha = 0.3) +
      #geom_text( aes(x=Var1,y=Var2,label = row_id),nudge_y=0.02,size=2) +
      theme( legend.position = 'none' )
  },
  final = m_df %>%
  {
    ggplot(.,aes(x=Var1,y=Var2,color=cluster) ) +
      geom_point(size=2 ) +
      lims( x=c(0,1.04), y=c(0,1.04) ) +
      #geom_text( aes(label = row_id),nudge_y=0.02,size=2) +
      theme( legend.position = 'none' )
  }
)
plot_grid( plts$original, plts$tsne, plts$assigned, plts$final, ncol = 2, nrow = 2, labels = c("1","2",
  print()
```

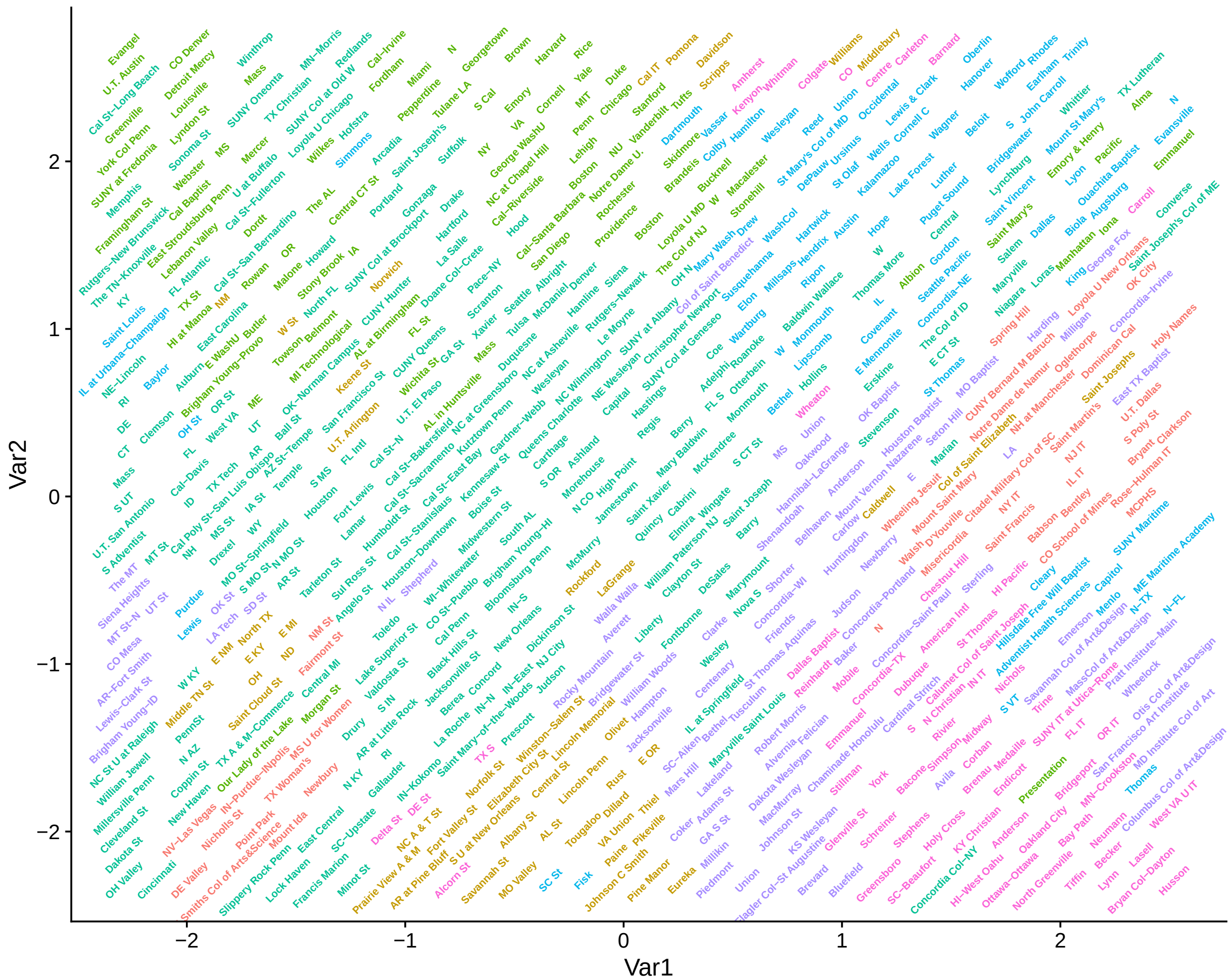


Perform Graph Alignment on t-SNE Coordinates

Now, we apply it to the college dataset t-SNE coordinates.

```
N_obs <- nrow( tsne_mat_hc_all )
grid <- expand.grid(1:floor(sqrt(N_obs)),1:ceiling(sqrt(N_obs))) %>% as.matrix() %>% {((.)-min(.))/diff
grid <- grid[1:N_obs,]

tsne_scaled <- tsne_mat_hc_all %>% {((.)-min(.))/diff(range(.))}
cost_matrix <- matrix(NA,N_obs,nrow(grid))
for( i in seq_len(nrow(cost_matrix))){
  for(j in seq_len(ncol(cost_matrix))){
    cost_matrix[i,j] <- sqrt( sum((tsne_scaled[i,] - grid[j,])^2) )
```

```
  }
}
cost_matrix = cost_matrix * (100000 / max(cost_matrix) )
rm( tsne_scaled )

px <- LinearAssignment( cost_matrix )

tsne_mat_hc_all %>%
  as_tibble() %>%
  mutate(
    College = names( cluster_id_all),
    cluster = factor( (cluster_id_all %% 7) + 1 )
  ) %>%
  slice( px ) %>%
  cbind( grid * diff(range(tsne_mat_hc_all)) + min(tsne_mat_hc_all) ) %>%
  {
    ggplot(.,aes( x = Var1, y = Var2 ) ) +
      geom_text(
        inherit.aes = FALSE,
        mapping     = aes( x = Var1, y = Var2 , label = College, color = cluster ),
        show.legend = FALSE,
        size        = 2,
        angle       = 45,
        fontface    = 'bold',
        check_overlap = TRUE
      )
  } %>%
  plot()
```



Conclusions

We do find some structure in the plot. And, the rotation of the axis to put Harvard University at the top-center helps us to interpret the axes and give meaning to that structure.

Notable Colleges

Clusters are colored with repeating colors and marked with repeating symbols, reflecting a limit of **ggplot2**. But each cluster should have an unique color-symbol combination.

Here are the t-SNE 2-D coordinates for some notable universities:

```
select_colleges <- c(
  '^OH St', '^MI-Ann Arbor', '^Purdue$', '^NU$','Harvard',
  'Yale', 'Princeton','^Penn$','^Cornell$','^Brown$',
  '^Howard$','Tuskegee','Hampton','Morehouse','Grambling',
  'Bethune-Cookman','Stanford','Johns Hopkins','Duke','Vanderbilt',
  'Rice','Wash.+St Louis','Notre Dame U\\.','^Pomona$','Harvey Mudd',
  'Swarthmore','MIT','Cal *IT','WI-Madison','IN-Bloomington',
  'Dartmouth',"Otis Col of Art&Design","San Francisco Art Institute",
  "Watkins Col of Art Design & Film","Rose-Hulman IT",
```

```r
  "Worcester Poly Institute","GA IT Campus","Davidson"
)

names( select_colleges ) <-
  c(
    "Ohio State","Michigan","Purdue","Northwestern",
    "Harvard","Yale","Princeton","Penn","Cornell","Brown",
    "Howard","Tuskegee","Hampton Inst","Morehouse","Grambling","Bethune-Cookman",
    "Stanford","Johns Hopkins","Duke","Vanderbilt","Rice","Wash.U.-St.L.",
    "Notre Dame","Pomona","Harvey Mudd","Swarthmore",
    "MIT","CalTech","Wisconsin","Indiana","Dartmouth",
    "Otis Col of Art&Design","San Francisco Art Institute","Watkins Col of Art Design & Film",
    "Rose-Hulman IT","Worcester Poly Institute","Georgia Tech","Davidson"
  )

rowid_select <- sapply( select_colleges, function(nm_regex) grep(nm_regex,tsne_df_all$College) )

sat_ugds_select <- DataSpec$student %>%
  slice( sapply( select_colleges, function(nm_regex) grep(nm_regex,college_names_student) ) ) %>%
  dplyr::select(1:2,UGDS,SAT_AVG,pctDisc1,pctDisc2,C150_4_POOLED_SUPP,CDR3,median_hh_inc_2005,pell_ever_
  mutate(
    UGDS = prettyNum( UGDS, big.mark = "," ),
    SAT_AVG = round(SAT_AVG),
    median_hh_inc_2005 = prettyNum(100*round(median_hh_inc_2005/100),big.mark=","),
    pctDisc_top2 = round(pctDisc1+pctDisc2),
    cluster = cluster_id_all[rowid_select]
  ) %>%
  dplyr::select(1:2,pctDisc_top2,everything(),-pctDisc1,-pctDisc2) %>%
  left_join(
    DataSpec$studentBF %>%
      dplyr::select(unitID,BF_discBreadth,BF_SAT_gt1400,BF_not1stgen,BF_fsend_5_2005,BF_CDR3),
    by = "unitID"
  )

tsne_select <- tsne_df_all %>%
  slice( rowid_select ) %$%
  set_rownames(as.matrix(select(.,Y1,Y2)),College) %>%
  round(1)
```

Group	College	Y1	Y2	SAT avg.	Cluster	Comments
Ivy League	Harvard	0	2.2	1501	16	
	Yale	0.1	2.1	1497	16	
	Penn	0	2.1	1442	16	
	Princeton	0.4	2.4	1495	20	
	Dartmouth	0.3	2.1	1446	18	
	Brown	0	2	1425	16	
	Cornell	-0.1	2	1422	16	
Big 10	Ohio State	-1.8	0.2	1289	32	
	Wisconsin	-1.7	0.1	1268	3	
	Purdue	-2	-0.6	1211	25	
	Indiana	-2	-0.5	1198	25	
	Michigan	-0.3	1.8	1352	16	is more like Ivies than Big10
	Northwestern	0	2	1458	16	is more like Ivies than Big10

Group	College	Y1	Y2	SAT avg.	Cluster	Comments
HBCUs	Howard	-1	0.9	1081	24	
	Tuskegee	-0.8	0.9	937	8	
	Hampton Inst	0	-0.7	990	5	
	Morehouse	-0.2	0	990	3	
	Grambling	-0.4	-1.6	863	1	
	Bethune-Cookman	-0.3	-1.6	812	1	
Arts Specialty	SF Art Inst	2.1	-1	1061	19	
	Otis C Art&Des	2.1	-0.8	1002	19	
	Watkins Art,Des,Film	2.1	-0.9	971	19	
Tech Specialty	Rose-Hullman	2	-0.2	1310	21	
	Georgia Tech	1.8	0.1	1352	21	
	WPI	1.9	-0.1	1256	21	
Others	Stanford	0.1	2.2	1466	16	
	MIT	0	2.1	1503	16	
	CalTech	0.2	2.5	1534	15	
	Johns Hopkins	-0.1	1.8	1418	16	
	Duke	0.1	2.2	1444	16	
	Vanderbilt	0.1	2.2	1475	16	
	Rice	0.1	2.1	1454	16	
	Wash.U.-St.L.	0	2.1	1474	16	
	Notre Dame	0	1.9	1450	16	
	Pomona	0.3	2.6	1454	15	
	Harvey Mudd	0.2	2.6	1483	15	
	Swarthmore	0.2	2.6	1442	15	
	Davidson	0.4	2.7	1353	15	

Interpretation of Quadrants

The combination of cluster locations and Bayes factors feature rays helps us assign meaning to each quadrant of the biplot.

Elite private & top-academic public, wealthy & smart

The vertical `Y2` axis is now almost perfectly aligned with the ray `pgt110K`, which is the ($\log_{10}$) Bayes factor capturing the prevalance of students from families with annual incomes greater than $110,000. All the Ivy League, "Ivy wannabes", and top-academic public universities (e.g., Cal-Berkeley, U. Michigan-Ann Arbor) are aligned along the positive vertical axis. That axis is almost perfectly countered by the downward-pointed ray `SATle800`, which is the Bayes factor capturing the prevalance of students with combined Verbal & Math SAT scores less than or equal to 800, i.e., the lowest tail of SAT scores.

Breadth versus specialization

The horizontal `Y1` axis isn't so readily interpretable. However, we see the ray `discBreadth`, whcih is the feature capturing the entropy (variety) in academic disciplines in which degrees are offered from the college, is pointing into the upper-left corner of the plot. So colleges aligned along this ray in the upper-right quadrant are the big public state universities that offer a broad range of degrees. On the other hand, the narrowly, highly specialized colleges appear in the lower-right quadrant of the plot.

Pell grants & high 3-yr credit default rates

The colleges in the lower-left quadrant are the colleges most strongly aligned with rays `pellever`, which captures prevalence of students having ever received a federal Pell grant, and `CDR3est`, which captures prevalence of students defaulting on student loans within 3 years of leaving the college.

More privates, but less elite

The upper-right quadrant is aligned with `SAT1400` (highest SAT students), `fsend5` (applied to many colleges), and `pgt48Kle75K` (mid-income families).

Summary

This was an exploratory analysis investigating structure in the U.S. Dept. of Education College Scorecard dataset.