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). 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.

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$studentBF %>%
  dplyr::select(
    c(-1, -(3:8)), -matches('[^4]_(WHITE|BLACK|ASIAN|OTHER|HISP|NRA|AIAN|UNKN)|2MOR|UNKN|NHPI|AIAN|BF_ma
    -matches('Challenge|_DEP_STAT_|notvet|le24y|OUTOFSTATE|prior|(^BF_[gl][et].+[0-9]+K$)|locale|FarWes
  ) %>%
  select_if( .predicate = function(x) any(x != x[[1]]) ) %>%
  filter( complete.cases(.) )
tsne_mat_all <- glmdata_all %>% select(-College) %>% as.matrix() %>% scale()
```

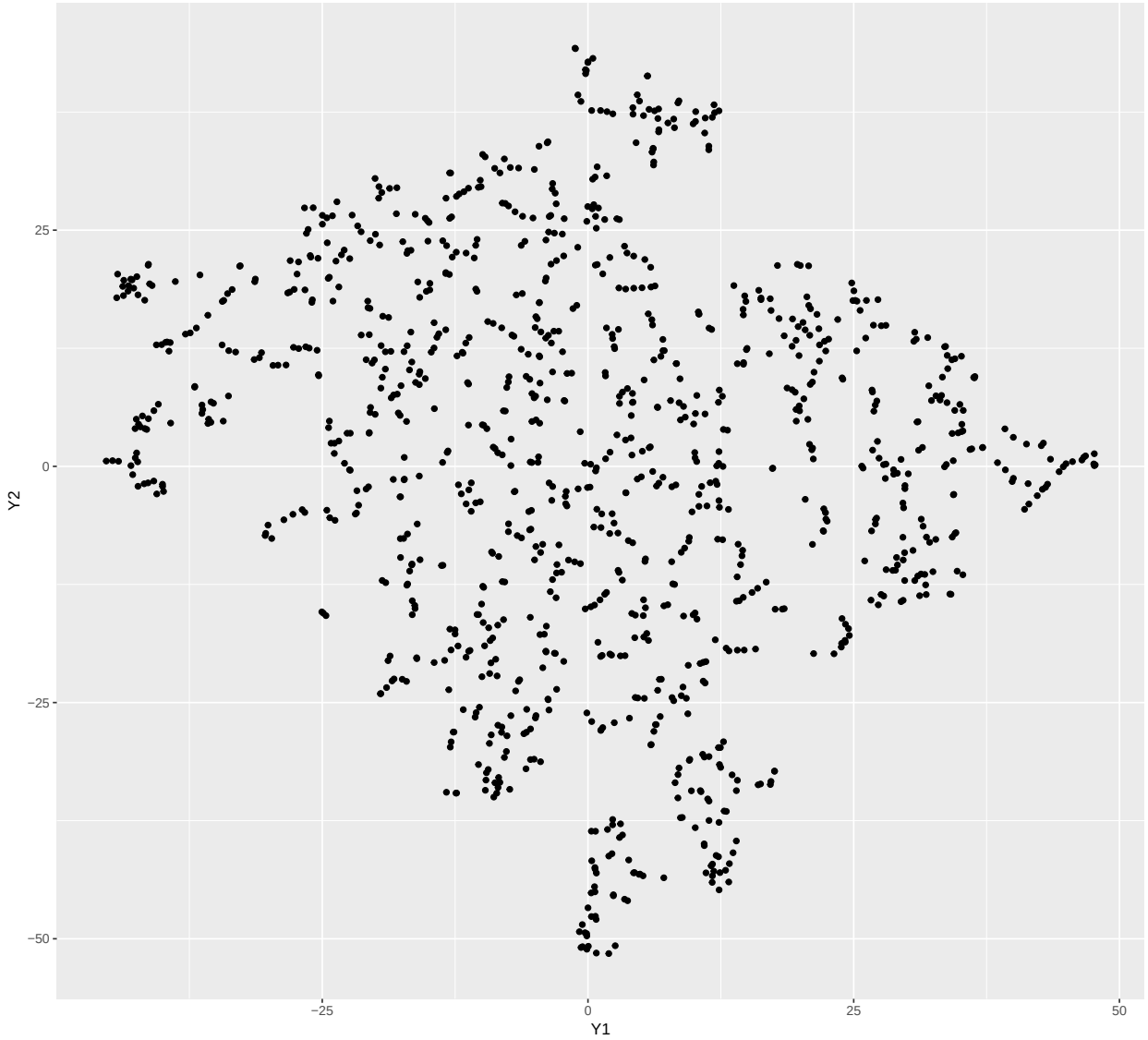
Perform t-Distributed Stochastic Neighbor Embedding (t-SNE)

Now, I'll map the data into a 2-D space. 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.

```
tsne_all <- Rtsne( tsne_mat_all, perplexity = 10, initial_dims = 12 )
```

```
tsne_all$Y %>%
  as_tibble() %>%
  setNames(c('Y1','Y2')) %>%
  {
    ggplot(.,aes(x=Y1,y=Y2)) +
      geom_point()
  } %>%
  print()
```



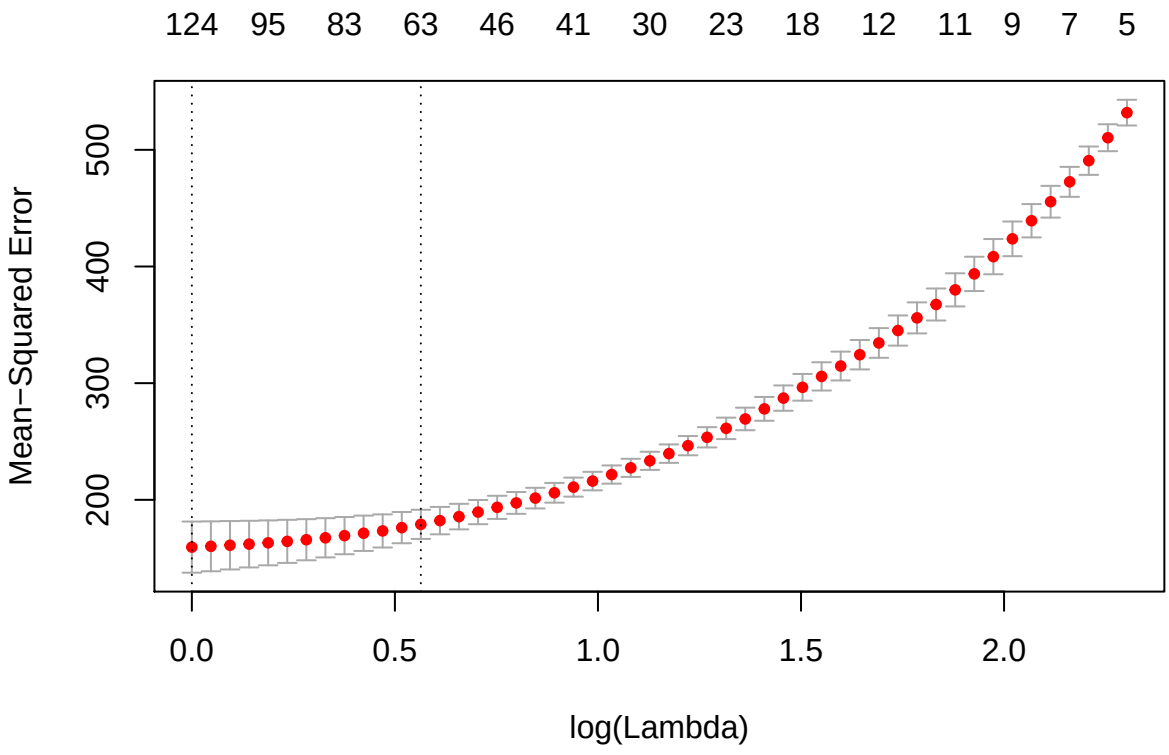
Find Underlying Dimension Driving 2-D Structure

I perform variable selection modeling of the 2-D t-SNE coordinates as responses vs. the original features from which the t-SNE coordinates were found. This way we'll have an approximate 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,
  y      = tsne_all$Y,
  family = 'mgaussian',
  lambda = exp(seq(log(1),log(10),length.out = 50))
)
plot( tsne_glmnet_all )
```

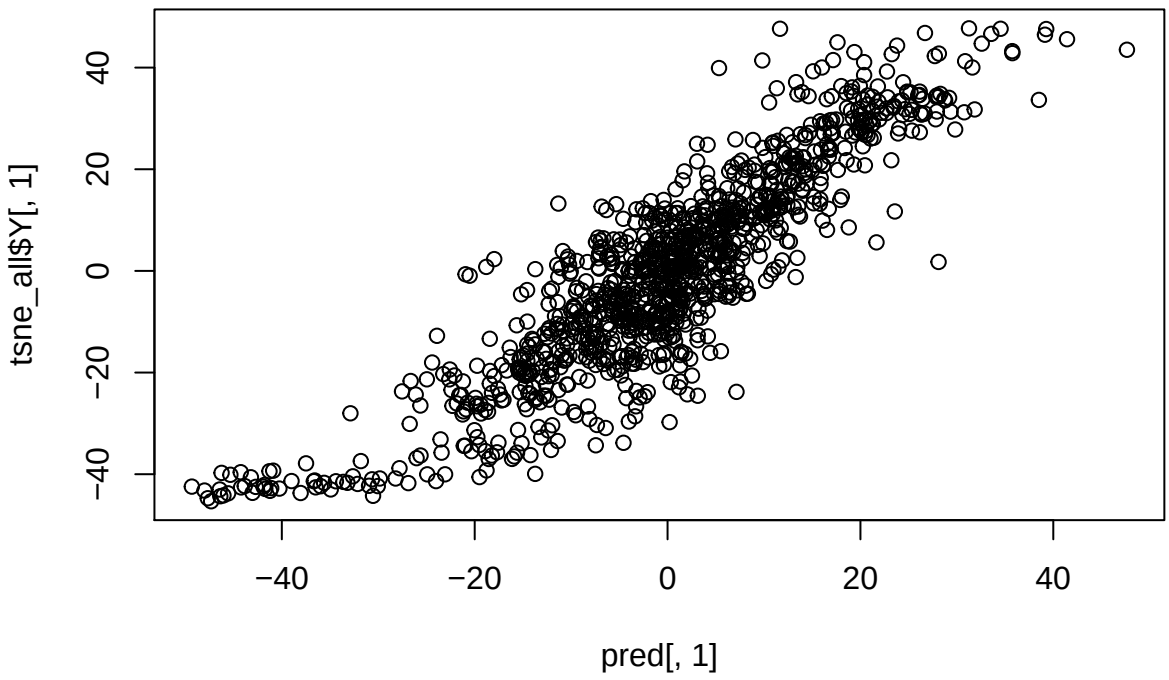


Check the Predictions

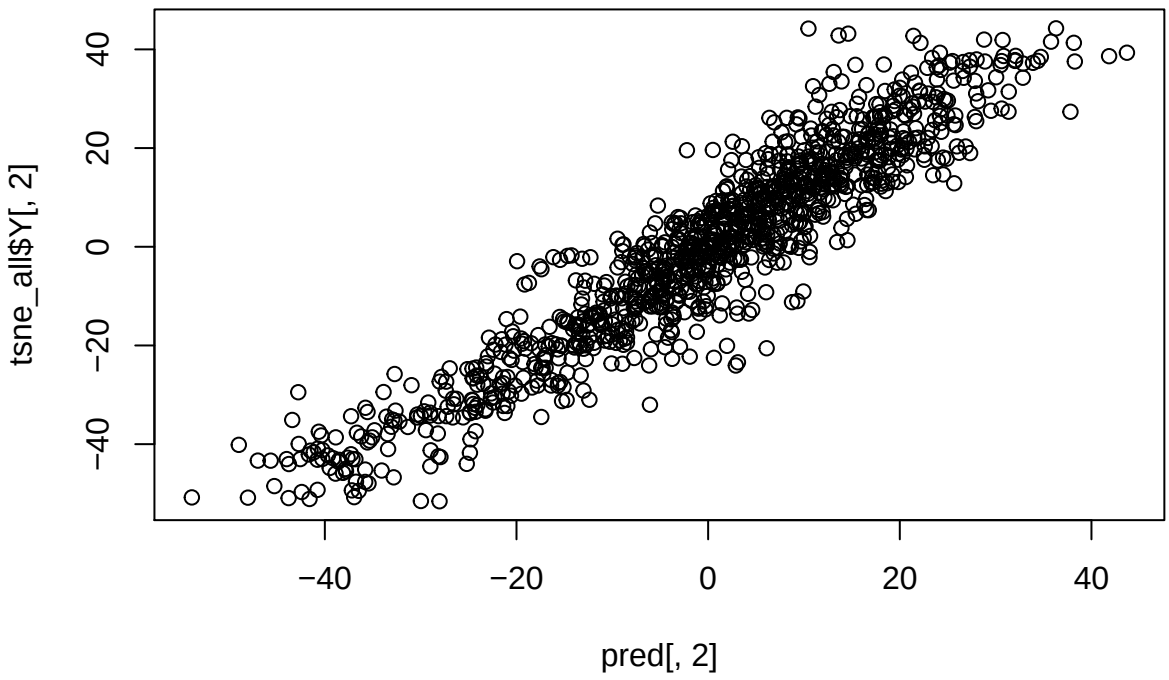
It's tricky to find a subset of features and their interactions that both describe the t-SNE coordinates well *and* do not suffer from extreme collinearity, which can make the validation error at low `lambda` explode when applying `glmnet`.

These predictions suck! (. . . at least for the moment.)

```
pred <- tsne_glmnet_all %>% predict( newx = mmat ) %>% drop()
plot(pred[,1],tsne_all$Y[,1])
```




```
plot(pred[,2],tsne_all$Y[,2])
```



Visualize the Colleges in 2-D

```
tsne_glmnet_coef_all <- tsne_glmnet_all %>% coef()
# tsne_glmnet_coef_all$y1[-1] %>%
# { (.)[abs((.)[,1])>0,1] } %>%
# { data_frame(Coefficient = names(.), value = round(.,2)) } %>%
#   print()
# tsne_glmnet_coef_all$y2[-1] %>%
# { (.)[abs((.)[,1])>0,1] } %>%
# { data_frame(Coefficient = names(.), value = round(.,2)) } %>%
#   print()

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)

tsne_coef_df_all %>% mutate(mag = sqrt(Y1^2+Y2^2)) %>% arrange(desc(mag)) %>% print(n = 30)
```

```
## # A tibble: 62 x 4
##                                        Coefficient          Y1          Y2
##                                              <chr>       <dbl>       <dbl>
##  1                                     discBreadth  6.89897456 -1.13156640
##  2           BF_SAT_gt800le1000:BF_MechanicRepair  4.52188190 -2.48386748
##  3                             BF_ForeignLanguages  2.62095846 -2.91061471
##  4                               BF_pell_ever_2005  1.88215417  3.29882472
##  5                                     BF_CDR3est  0.43262372  2.58360721
##  6                                  BF_SAT_gt1400  0.06054633 -2.39401630
##  7                                BF_fsend_5_2005 -0.70342780 -1.99909437
##  8                                  BF_AreaEthnic  1.44381743 -1.52434117
##  9                             BF_PhysicalSciences  1.51208892 -1.31990204
## 10                                 BF_p_gt48Kle75K -0.74981941 -1.80888781
## 11                               BF_FamilyConsumer  1.45462251  0.23151819
## 12                                   BF_SAT_le800 -0.43820502  1.34226359
## 13                                    BF_veteran  0.80161435  0.97244618
## 14     BF_fsend_5_2005:BF_EngineeringTechnologies -0.83545833  0.51491427
## 15                         BF_PhilosophyReligious  0.51194248 -0.83280940
## 16                                    BF_p_gt110K  0.40675907 -0.81816608
## 17         BF_EngineeringTechnologies:discBreadth  0.87023773  0.13173864
## 18                                  BF_gt24yrsold  0.24117630  0.79813849
## 19                    BF_EngineeringTechnologies  0.77580617  0.08092055
## 20            BF_fsend_1_2005:BF_ForeignLanguages  0.43713167  0.62031961
## 21              BF_Education:BF_TheologyReligious -0.69274807 -0.14028800
## 22                      BF_AgricultureAgriculture  0.64960762  0.14458942
## 23                              BF_SocialSciences  0.54328015 -0.28940255
## 24                         BF_ArchitectureRelated  0.60956447 -0.06092011
## 25                                 BF_Engineering  0.54763791 -0.26239126
## 26                            BF_NaturalResources  0.42558749 -0.41776865
## 27                     BF_fsend_5_2005:BF_CDR3est -0.42513556  0.41023778
## 28                             BF_HomelandSecurity  0.37743022  0.43620879
## 29                          BF_MathematicsStatistics  0.45686188 -0.26410392
## 30 BF_EngineeringTechnologies:BF_ForeignLanguages  0.45887320  0.17808015
## # ... with 32 more rows, and 1 more variables: mag <dbl>
```

```
# tsne_coef_df %>%
# {
#   ggplot(., aes(x=Y1,y=Y2,label=Coefficient)) +
#     geom_point() +
#     geom_text( check_overlap = TRUE )
# } %>%
#   print()

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)")


college_names <- glmdata_all %$%
  College %>%
  { gsub('^[0-9_]+','',. ) } %>%
  { gsub('Northwestern University','NU',.) } %>%
  { gsub('University of Notre Dame','Notre Dame U.',.) } %>%
  { gsub('Cornell University','Cornell U.',.) } %>%
  { gsub('California','Cal',. ) } %>%
  { gsub('Mass.+Inst.+Tech.+','MIT',. ) } %>%
  { gsub('(Mass|Penn|Wash)[^ ]+ *','\\1',.) } %>%
  { gsub('Polytechnic','Poly',. ) } %>%
  { gsub('Institute of Tech[^ ]+','IT',. ) } %>%
  { gsub('Tech.+Inst.+','Tech',. ) } %>%
  { gsub('State','St',. ) } %>%
  { gsub('University','U',. ) } %>%
  { gsub('(U of )|( U$)','',. ) } %>%
  { 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],.) }
}

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()

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(1.7*scale(.)),.vars=vars(Y1,Y2))
```

Show Biplot for Structure Interpretaion

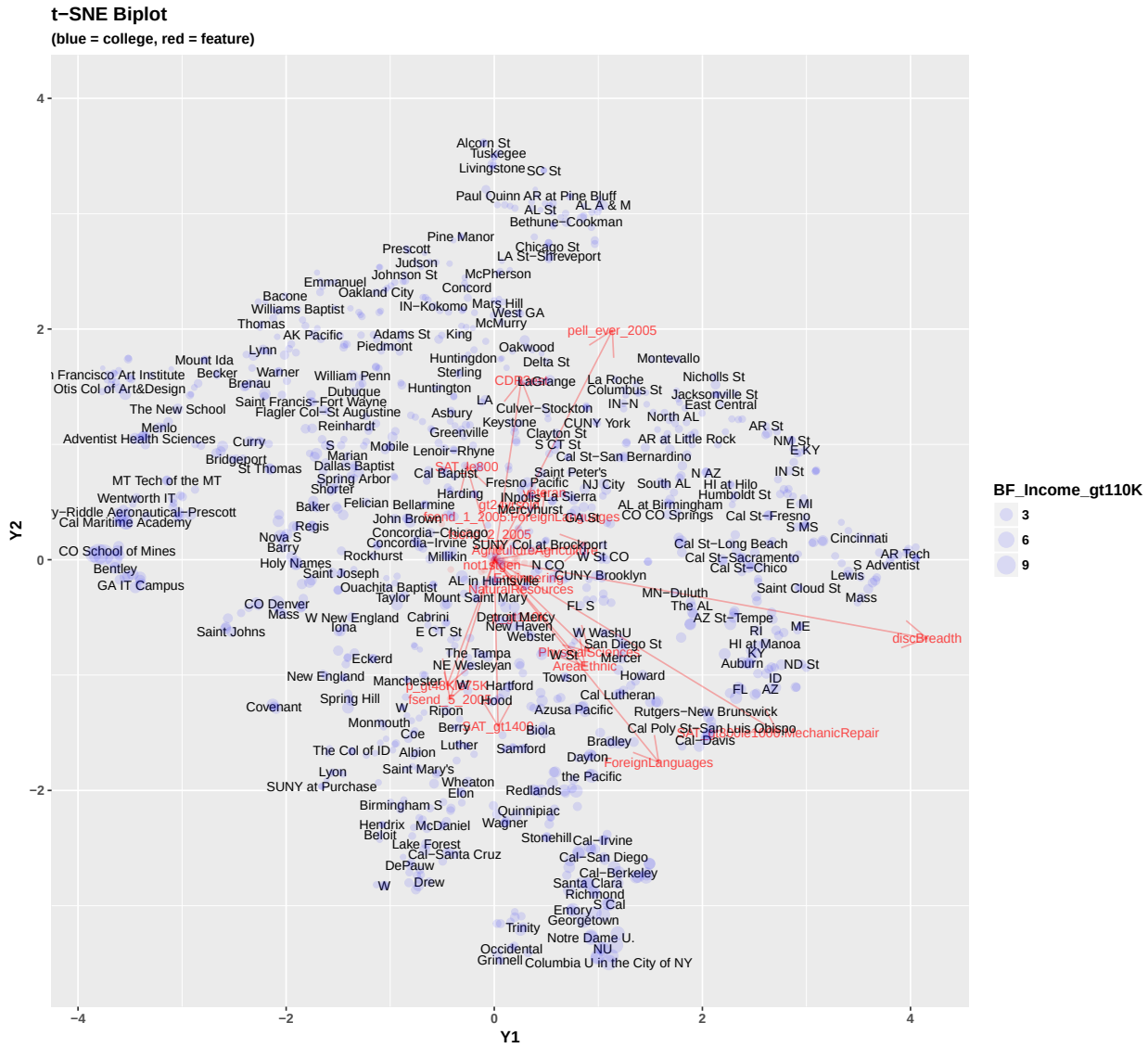
we can overlay the feature dimensions on the college scatterplot in the 2-D t-SNE coordinate space. This allows us to more easily interpret the structure we're seeing. However, some of the interaction terms, in particular, are tricky to interpret because they have a positive value for a college if both 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.

```
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
y2_min <- -3.5
tsne_coef_df_all %>%
  mutate(
    mag = sqrt(Y1^2 + Y2^2),
    Y2 = pmax(y2_min,Y2*f_mult),
    Y1 = Y1*f_mult,
    Coefficient = gsub('\\([^)]+\\)','',gsub('BF_','',Coefficient))
  ) %>%
  {
    ggplot(., aes( x = Y1, y = Y2 ) ) +
      geom_point( color = 'red', alpha = 0.1 ) +
      geom_text(
        aes( label = Coefficient),
        color = 'red',
        alpha = 0.7,
        size = 3,
        check_overlap = TRUE
      ) +
      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_df_all,
        aes( x=Y1, y=Y2, label=College ),
        mapping=,
        color = 'black',
        size=3,
        check_overlap = TRUE
      )  +
      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()
```

```
## Warning: Removed 37 rows containing missing values (geom_text).
```

```
## Warning: Removed 37 rows containing missing values (geom_point).
```




```
select_colleges <- c(
  '^OH St', '^MI-Ann Arbor', '^Purdue$', '^NU$',
  'Harvard', 'Yale', 'Princeton','^Penn$','^Cornell U\\.$','^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'
)
tsne_select <- tsne_df_all %>%
```

```
  slice( sapply( select_colleges, function(nm_regex) grep(nm_regex,(.)$College) ) ) %$%
  set_rownames(as.matrix(select(.,Y1,Y2)),College) %>%
  round(1)
```

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

- **Big 10**
  - Ohio State: 2.5, -1
  - Michigan: 1.5, -2.7
  - Purdue: 3.7, -0.2
  - Northwestern: 1, -3.4
- **Ivy League**
  - Harvard: 1, -3.7
  - Yale: 1, -3.5
  - Princeton: 0.2, -4.1
  - Penn: 1.2, -3.3
  - Cornell: 1.2, -3.2
  - Brown: 1, -3.4
- **HBCUs**
  - Howard: 1.4, -1
  - Tuskegee: 0, 3.5
  - Hampton Institute: 0.2, -0.9
  - Morehouse: -0.3, -1.1
  - Grambling: 0.7, 3
  - Bethune-Cookman: 0.7, 2.9
- **Others**
  - Stanford: 1.1, -3.5
  - Johns Hopkins: 1, -3.1
  - Duke: 1, -3.4
  - Vanderbilt: 1, -3.5
  - Rice: 1, -3.5
  - Washington U.-St. Louis: 1.1, -3.4
  - Notre Dame: 0.9, -3.3
  - Pomona: -0.1, -4.2
  - Harvey Mudd: 0.1, -4.2
  - Swarthmore: 0, -4.2
  - MIT: 1.1, -3.6
  - CalTech: 0.2, -4.2

Perform Hierarchical Clustering

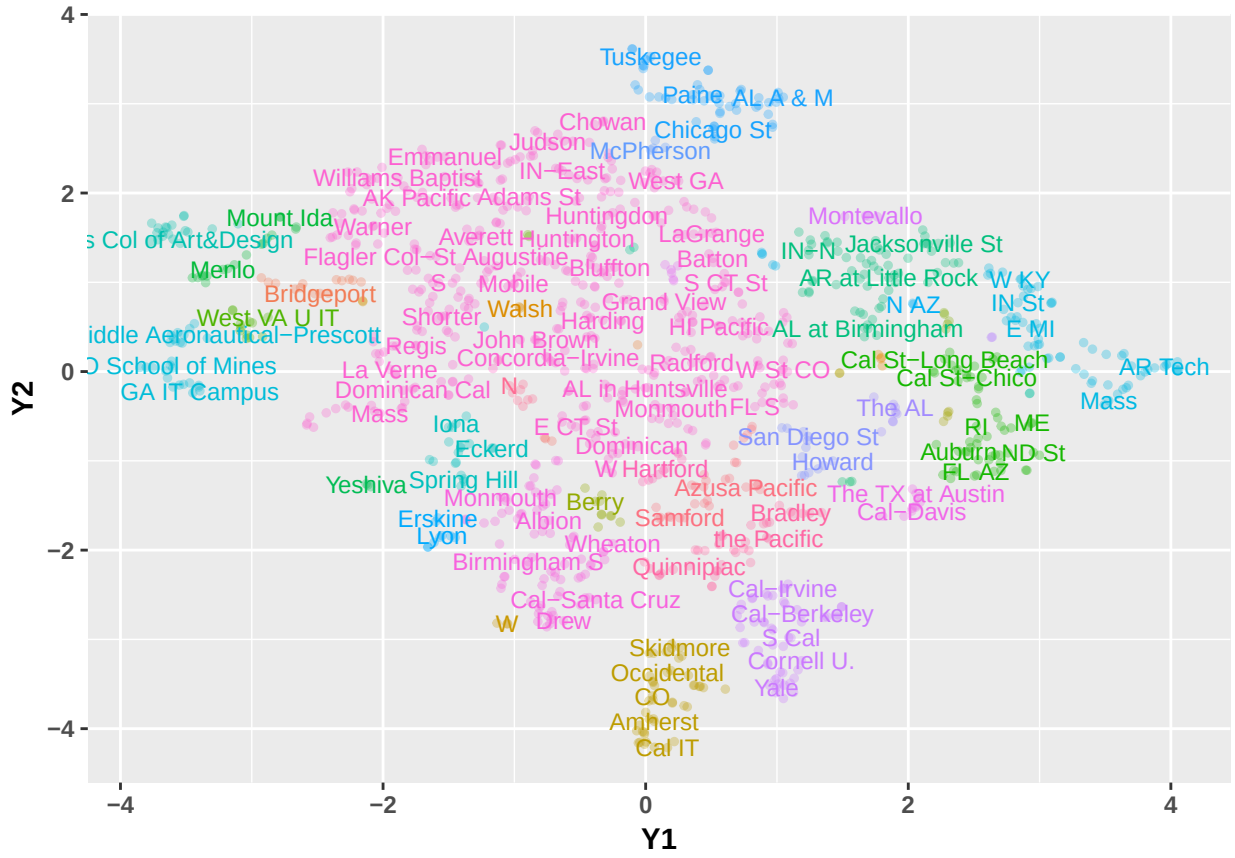
Now, I perform cluster analysis. Hierarchical clustering is a quick way to identify clusters in the 2-D t-SNE space. We can then color the clusterings in a scatterplot to more easily visualize the structure.

```
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 <- 45
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) )
```

```
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()
```



Show Biplot with Cluster Coloring

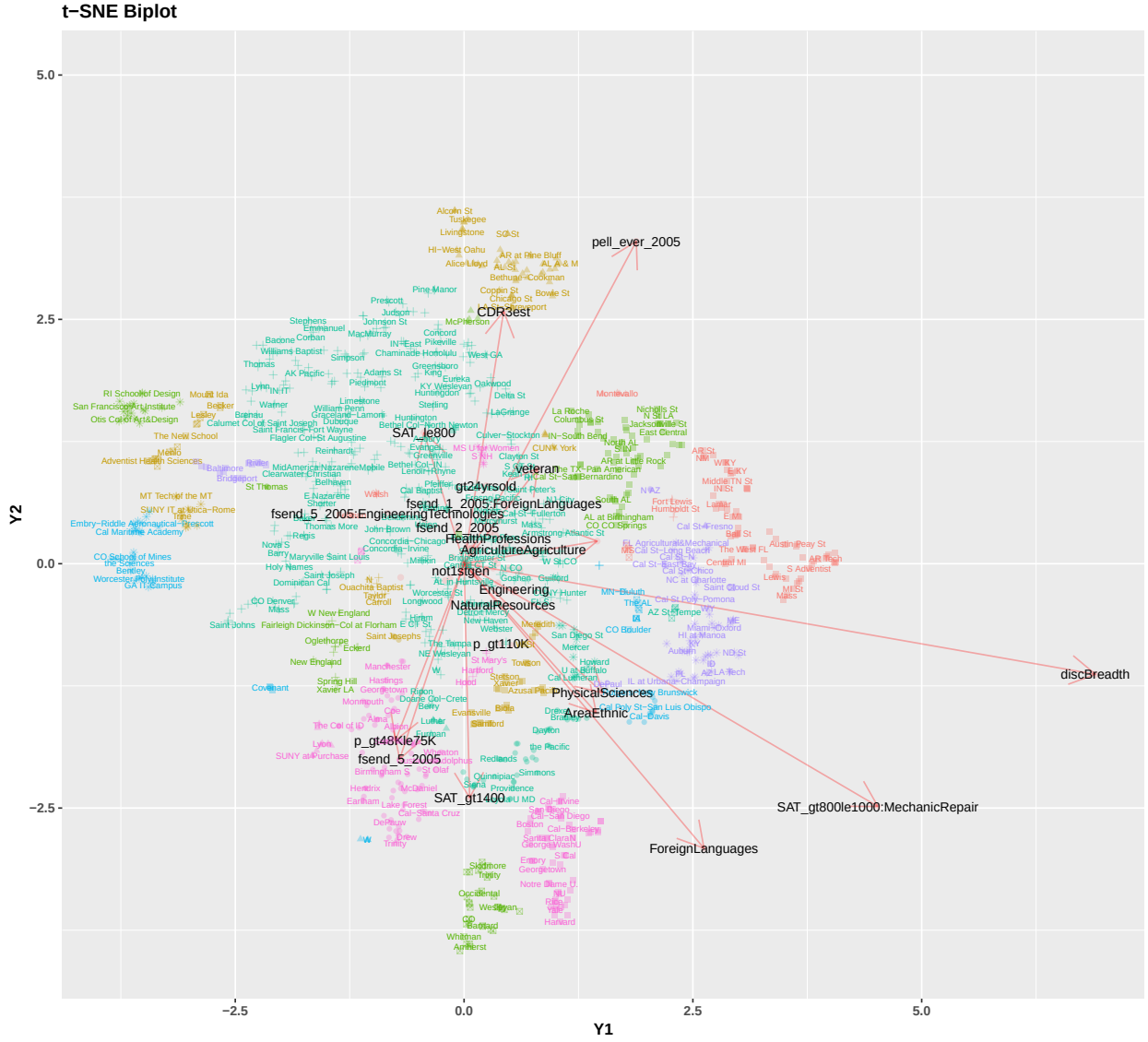
Finally, we can overlay the feature dimensions on the 2-D

```
cluster_id_all <- cutree( hc_all, k = n_cluster )
y2_min <- -4
tsne_coef_df_all %>%
  mutate(
```

```
    mag = sqrt(Y1^2 + Y2^2) ,
    Y2 = pmax(y2_min,Y2),
    Coefficient = gsub('\\([^)]+\\)','',gsub('BF_','',Coefficient))
  ) %>%
  {
    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( "t-SNE Biplot" ) +
      theme( text = element_text( face = 'bold' ) ) +
      scale_y_continuous(limits = c(y2_min,5))
  } %>%
  print()
```

```
## Warning: Removed 11 rows containing missing values (geom_text).
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



Conclusions

TBD

Summary

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