**HUDK 5053 Feature Engineering Studio**

**- Deliverable 4: Visualization (October 13, 2016) –**

M.S. Learning Analytics

Sunmin Lee

*Research question: Build a prediction model to find out key features of drop-out students in K-12.*

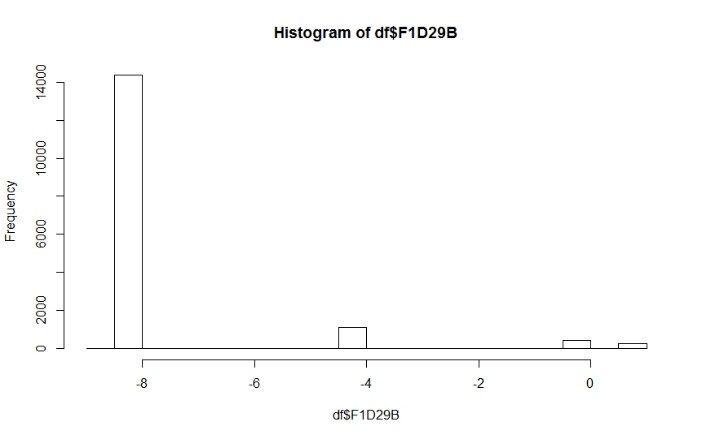
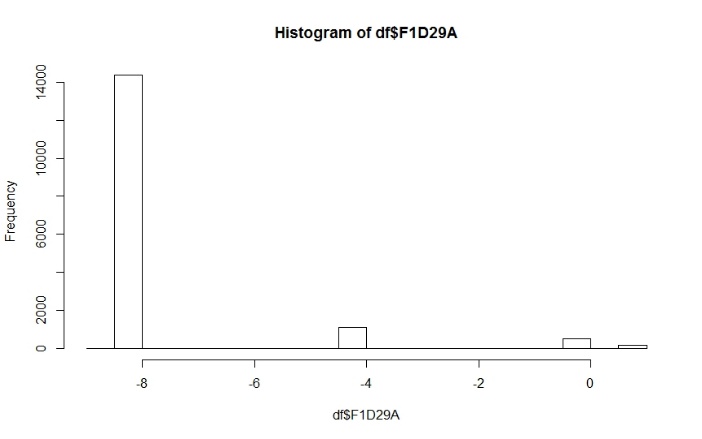
**1. Data cleaning**

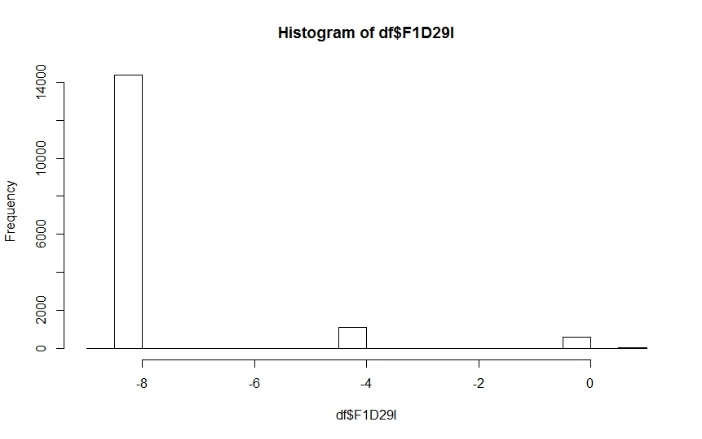
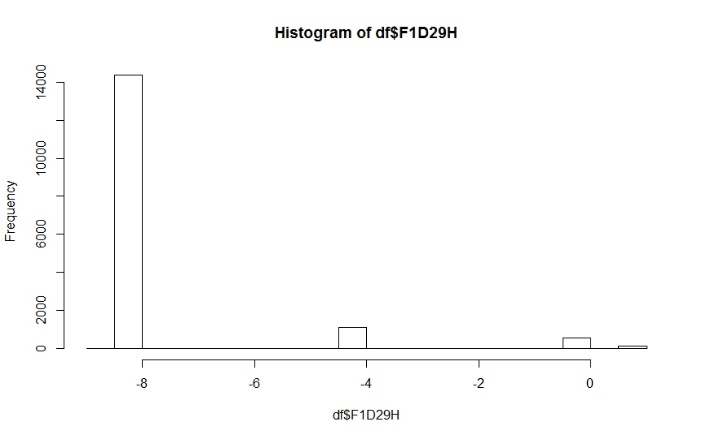
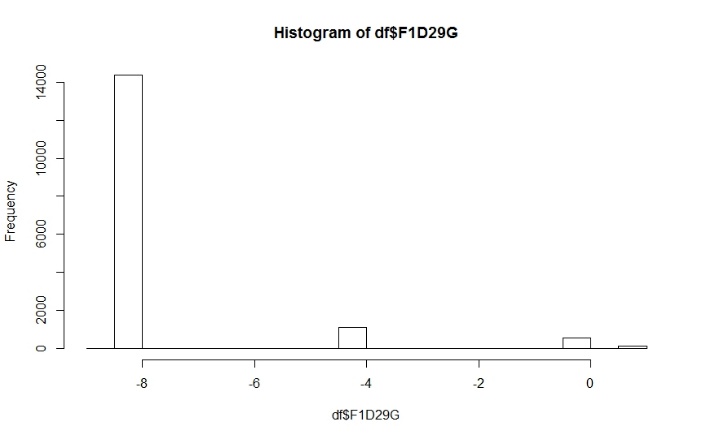
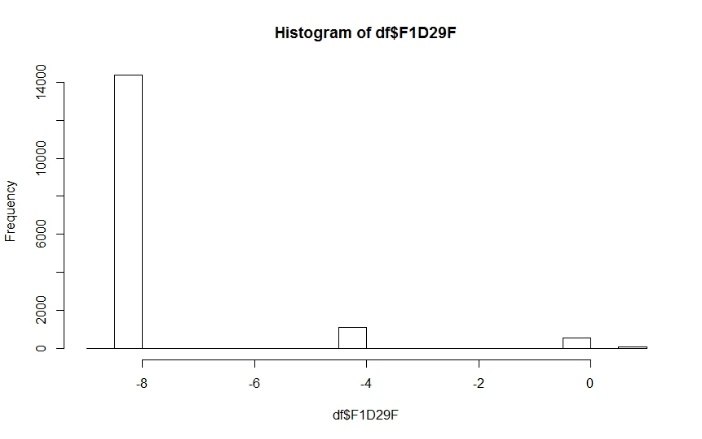
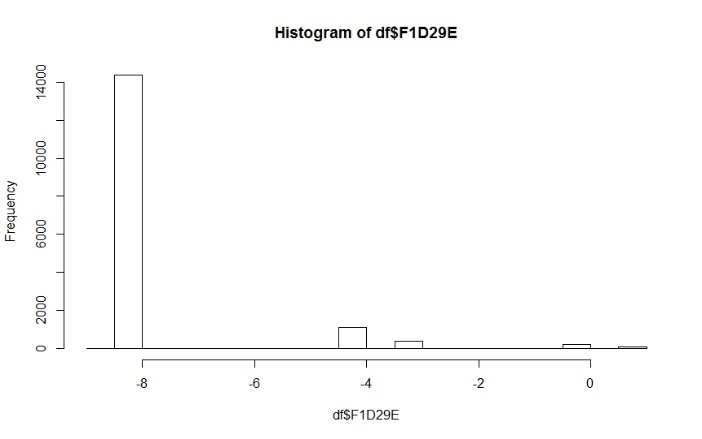
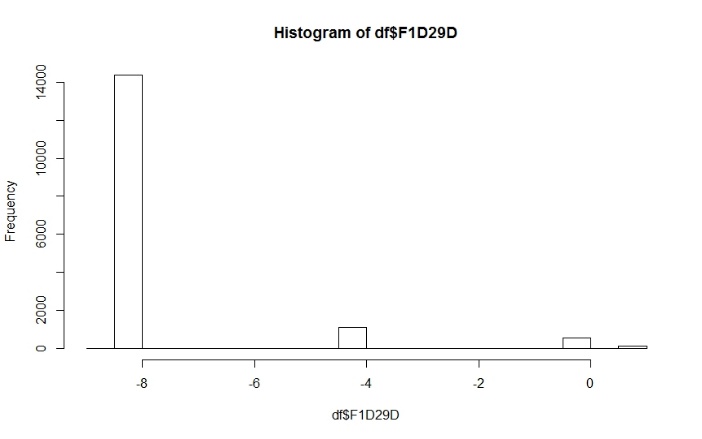
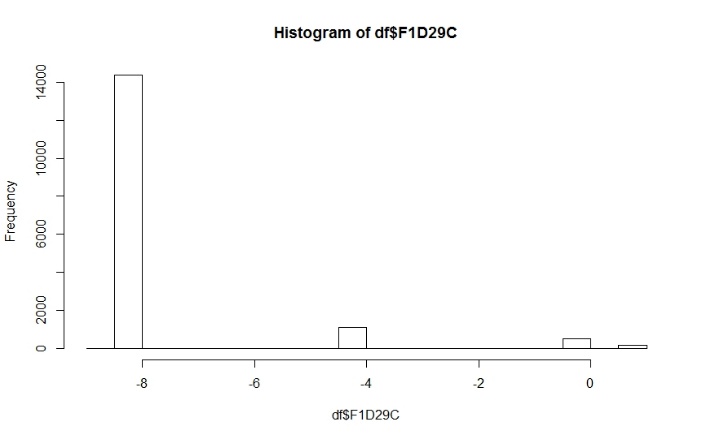
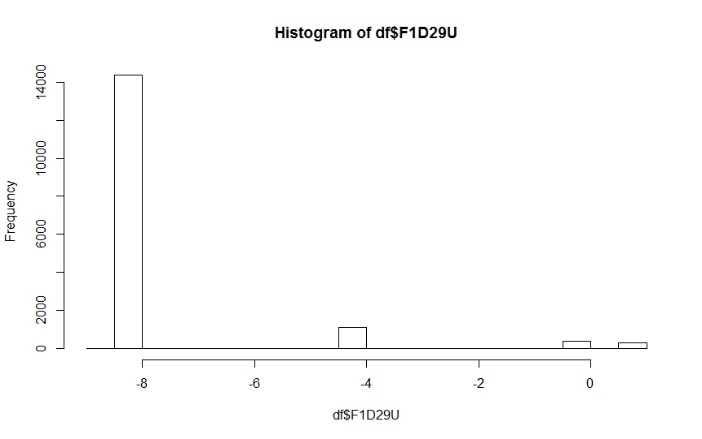
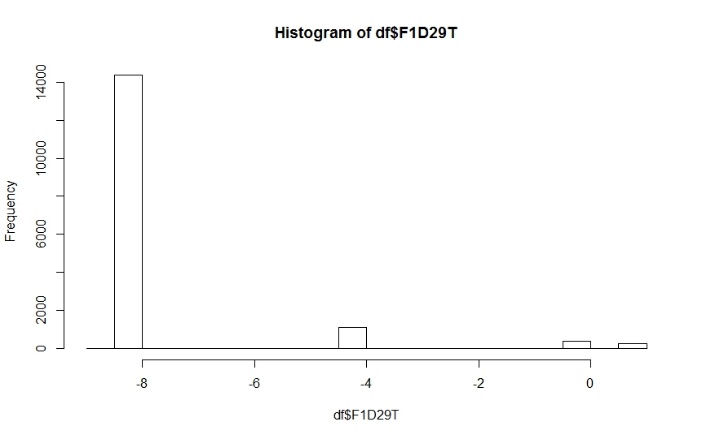
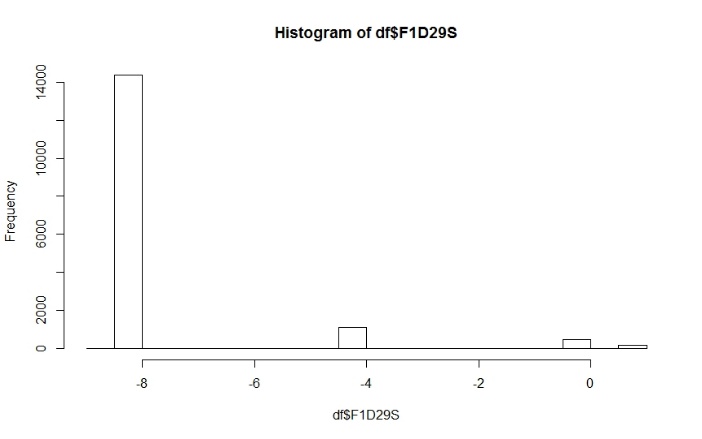
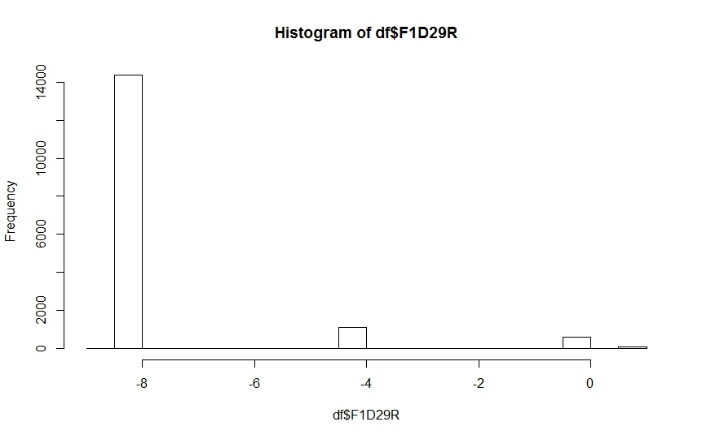
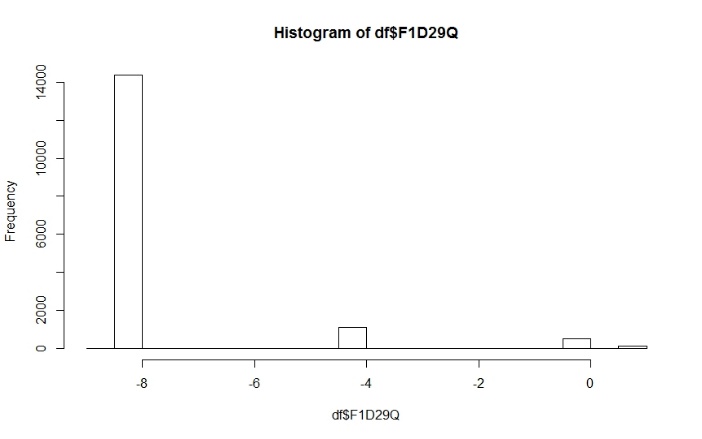
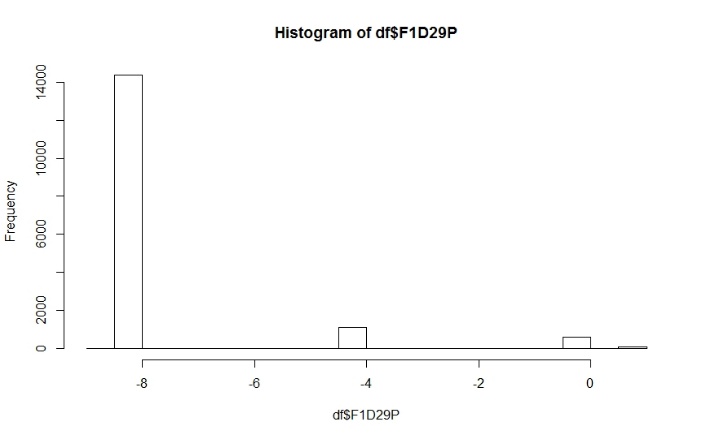
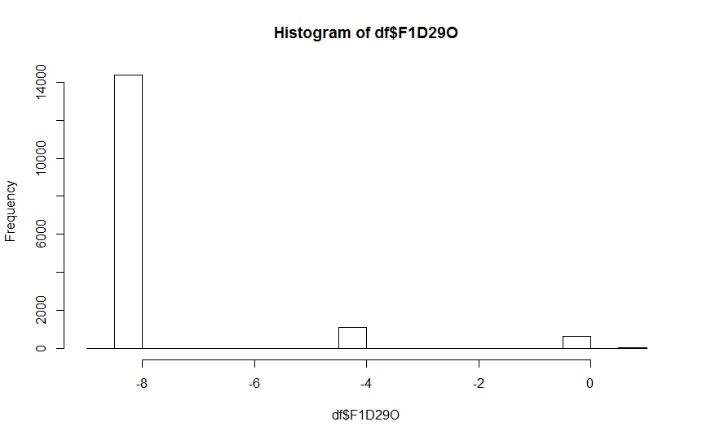
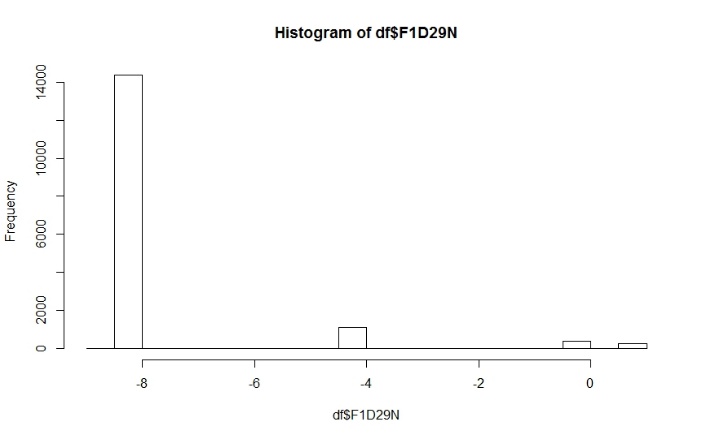
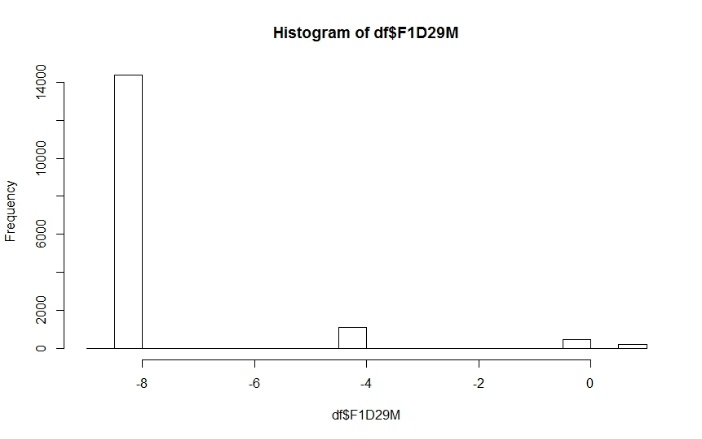
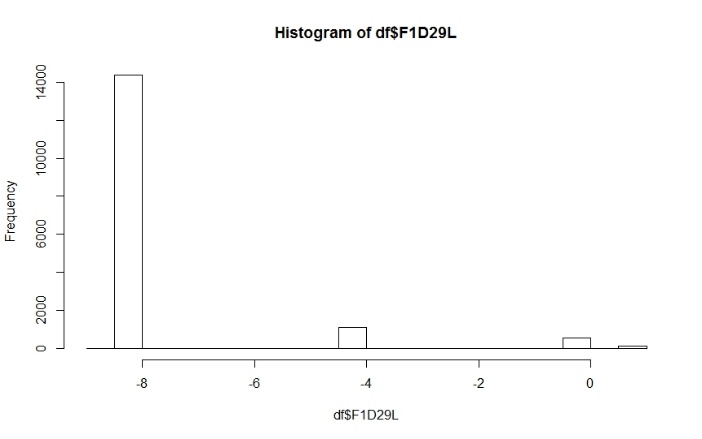
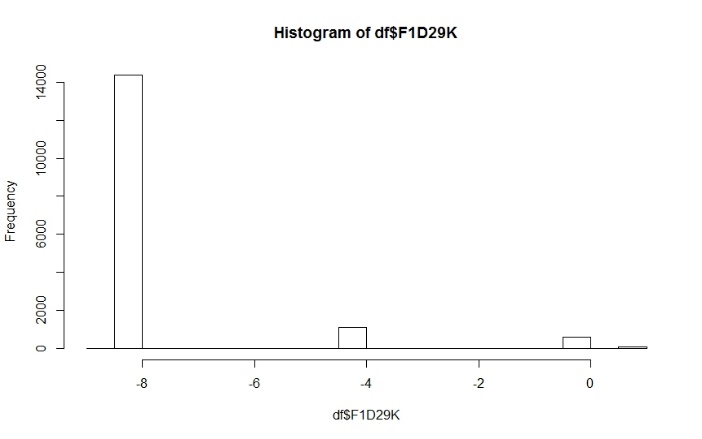
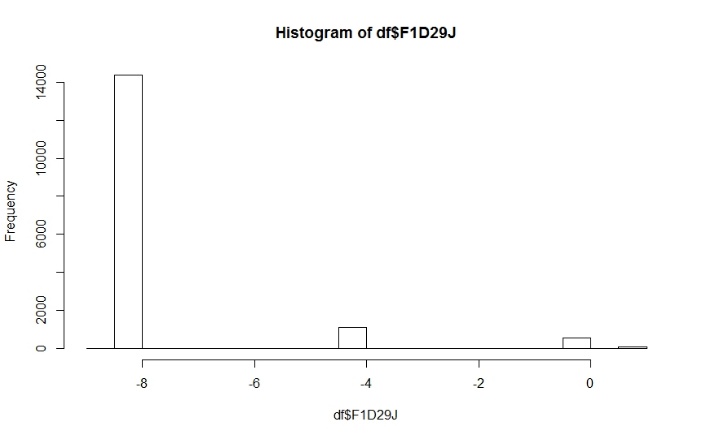
In assignment 3, I decided to use the Education Longitudinal Study of 2002 (ELS: 2002) provided from National Center for Education Statistics (NCES). Within the ELS 2002 study surveys, I pulled out the questions related to “drop-out students” and in total, 4,012 variables with 16,197 observations were found. However, since this file was too big (167MB) to run analysis in R with my computer capacity and I didn’t need all variables, I decided to clean the data with M.S. Excel for better efficiency. Eventually, I kept 21 variables (F1D29A - F1D29U) which were answering the Question #29 in the survey, “Here are some reasons other people have given for leaving school. Which of these would you say applied to you?”, as I was interested in finding out the key features of drop-out students.

**2. Visualization 1 – Histogram**

In a first step after cleaning data, it was difficult to figure out how many data of those 16,197 observations will be useful for my analysis. Hence, I started to make a histogram for every variable to understand the distribution of my data. As Figure 1 shows, most of the data (around 96%) were missing or not available, unfortunately. This is because the Question #29 applies only to those who actually dropped out and the response rate is not expected to be very high. It was also true for all 21 variables and made me decide to recode the missing values.

*Figure 1: Histogram*



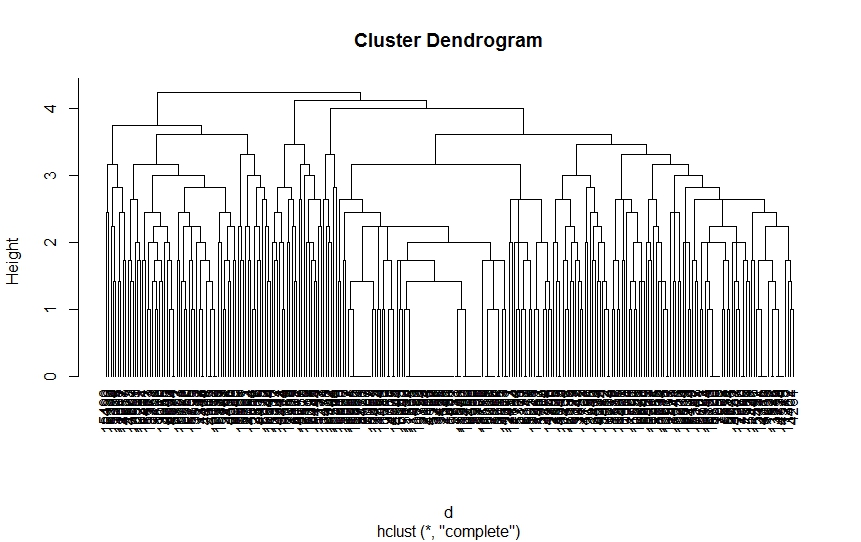
**3. Recoding missing values**

To save time in re-coding missing data (and I didn’t want to repeat the R code for all 21 variables), I used the M.S. Excel again to change all missing values into “999”. According to the original dataset, missing or N/A (or other reasons that resulted in non-usable data) were coded from “-9” to “-1”. For the convenience (and change a large amount of data at once), I changed all these values into “999” in M.S. Excel and coded as “NA” in R for every variable. In order to check whether it was coded correctly, I generated a histogram to see if they have other values than “0” and “1”. In the end, this dataset contained only binary data (i.e. “yes” or “no”) with 282 observations.

**4. Visualization 2 – Cluster dendrogram**

Since I wanted to identify the relationship between drop-out students, I was initially thinking of making a scatter plot. However, unfortunately, I had only binary data and it was difficult to make scatter plot so as to check the relationship. Hence, I decided to make a cluster dendrogram, although I knew it will be challenging due to massive observations (see Figure 2). Yet, it was a good chance for me to understand the distribution of this dataset since I found that there exist four major groups.

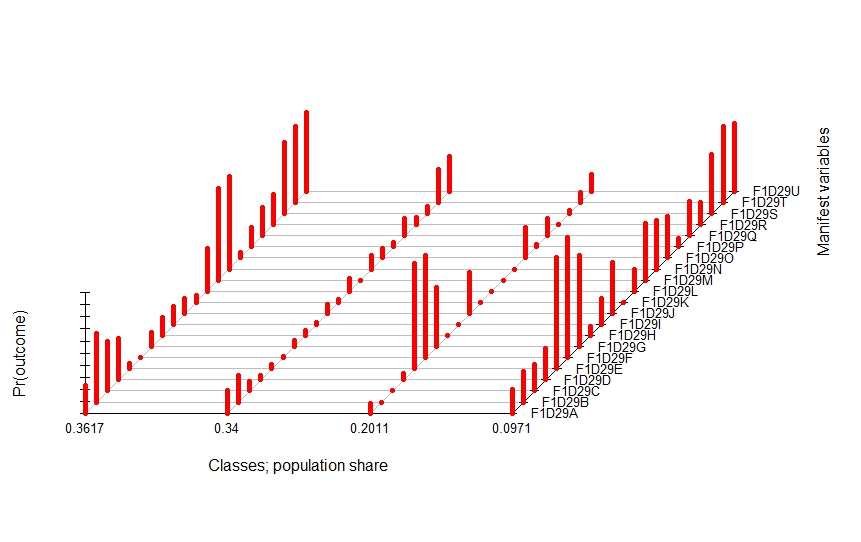
*Figure 2: Dendrogram*



**5. Visualization 3 – Latent Class Analysis**

Although I couldn’t figure out which variables belong to each group, there was a way to find this answer; using latent class analysis. Hence, I run the LCA model in R (“Library-poLCA”) with the cleaned data. Since I already knew there are four big clusters in my dataset, I intentionally set up into four classes regardless of AIC and BIC and plotted the results (see Figure 3). According to the result, 34% of the dataset was in class 1, 36% in class 2, 20% in class 3, and 9.7% in class 4. Also, we can clearly see which variables belong to specific classes.

*Figure 3: Latent Class Analysis*



**6. Further step**

I didn’t go further for next step in this assignment. But with this result, I assume that we can generate a heatmap by groups. We can aggregate variables into a group and assume the possible drop-out reason (i.e. latent). Otherwise, we can line up all the 21 variables and generate a heatmap with “yes” or “no”.

**7. Obstacles that we discussed**

Now I figured out which variables are related to each other and make a group accordingly. However, I am not sure how I can match this to the prediction model. My initial research question was to make a prediction model to figure out key features for drop-out students. I am confused with which variable should I use for predicted variable (for instance, student learning achievement) and I assume there will be many other factors that affect my question besides selected variables. Also, having only binary data gives me many limitations in analyzing.