Assignment II

Matthew Adair

4/2/2018

Contents

Introduction

The intent of this assignment is to find insights into the admissions data through hierarchial clustering and k-means clustering. I am particularly intereseted in the statistics for each cluster in regards to MeritAward, AcdmcIndex, ACTComp, GPAREcalc, Decision, EventParticipation and CampusVisits to see how widely the clusters may vary or not vary from one another. The admissions data that I used began with 24 variables, nine of which were categorical variables and the rest were quantitative variables. More specifically, of those fifteen quantitative variables, only six of the variables are numeric which are the variables EventParticipation, CampusVisits, ClassSize, GPARecalc, ACTComp, and MeritAward, while the remaining nine quantitative variables are factored variables.

Cleaning

I began with the cleaning process by eleminating the qualtitative varaibles that I was not able to convert into numeric types without obstructing data analysis or misrepresenting the data. So I removed the variables X, which seemed to be an additional column that was randomly genertaed when reading in the admissions data, and ID to begin with. The reason why I removed ID was beacuase that variable did not add any value to the analysis since it's whole purpose was to uniquely identify each observation, or student in this case. I further removed the variables Sex, Race, Sport, AcdmcInt1, PrmntGeom, and CitizenshipStatus. The reason why I removed these variables is due to the nature of these varaibles and them not being able to convert into a logicla way into numeric data tha would not skew the analysis and misrepresent the data.

I then converted and revalued categorical variables. I began with Inquired, meaning if a student inquired about anything in regards to the university before they applied, and PermntCountry, which holds data about where an accepted student is currently living, into binary numeric variables. The reasons why I converted those two variables into binary numeric variables were because they were already binary factor variables making the conversion logical and beneficial to the analysis because I would be able to retain more varaibles and retain some statistical power. Furthermore, I converted DecisionPlan into a numeric variable by weighing the Early Decision plans more than Early Action plans which were in turn weighed more than Regular Decision plans. Additionally, I converted SportRating it into a factored numeric variable placing "Blue Chips" equal to three, "Franchise" equal to two, "Varsity" equal to one, and "None" as zero. After making all of the

minor tweaks to the data I converted all of the variables that were factors into numeric data so that they could be used for the various forms of cluster analysis that I will perform.

Hierarchial Clustering

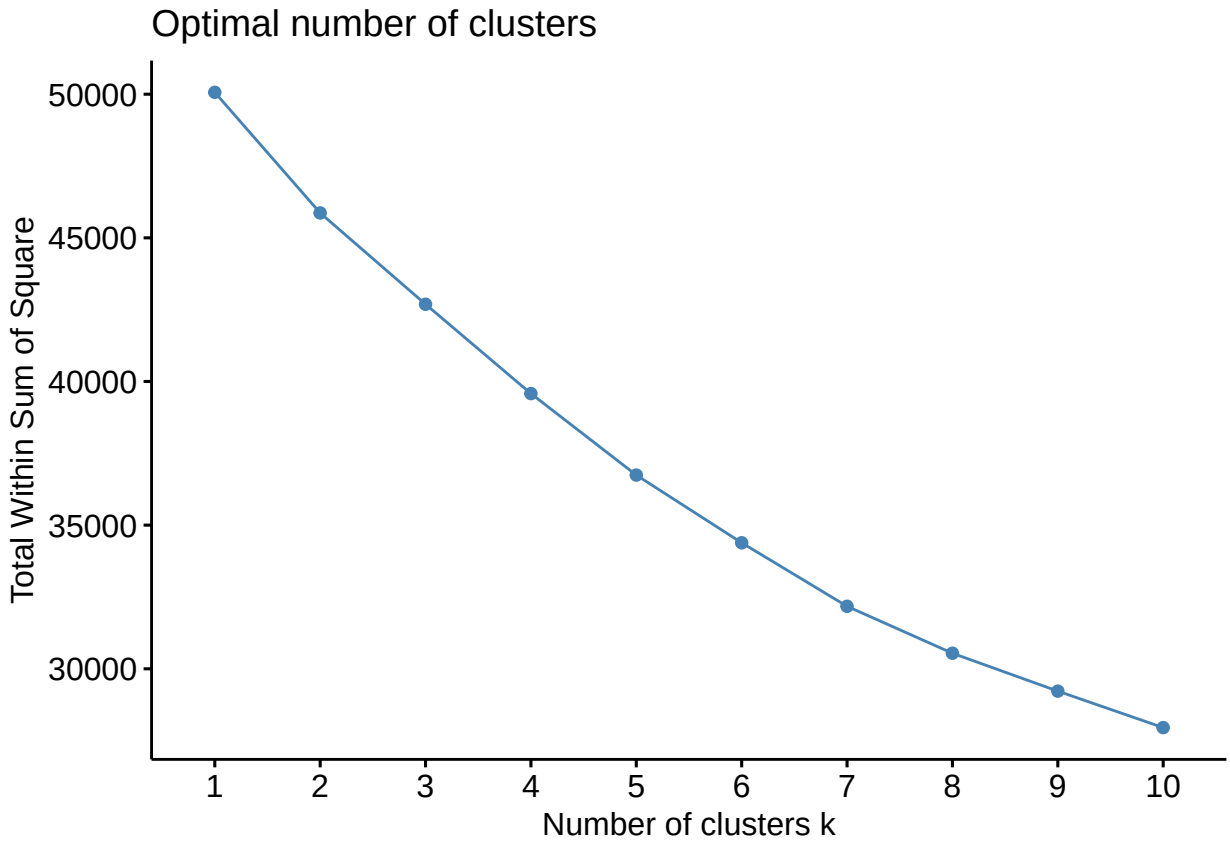
I began with creating different hierarchial models based off of complete, average, and single linkage. The data I first used was not scaled and when viewed, the heights of the various dendrograms did not make a lot of sense and so I switched over to a scaled admissions dataset which created a dataset where the variables could now be compared in a way that is readable. I split up the hierearchial clustering between two algorithms, the first one of which is Agglomerative clustering or AGNES for short, and the other was Divisive hierarchial clustering which is also reffered to as DIANA. Furthermore, I applied four different methods to the AGNES models, which were average linkage, single linkage, complete linkage, and Ward's minimum variance method. In order to decide which method was best I calculated the agglomerate coefficient for each method as shown below. The values closer to one means the stronger the clustering strucuture is. As seen from the results below, the method that produced the strongest clustering structure was Ward's method. I also performed a hierarchical model based off of correlation distance as well as centroid linkage. Those models however, I did not use as my final model for AGNES hierarchical methods.

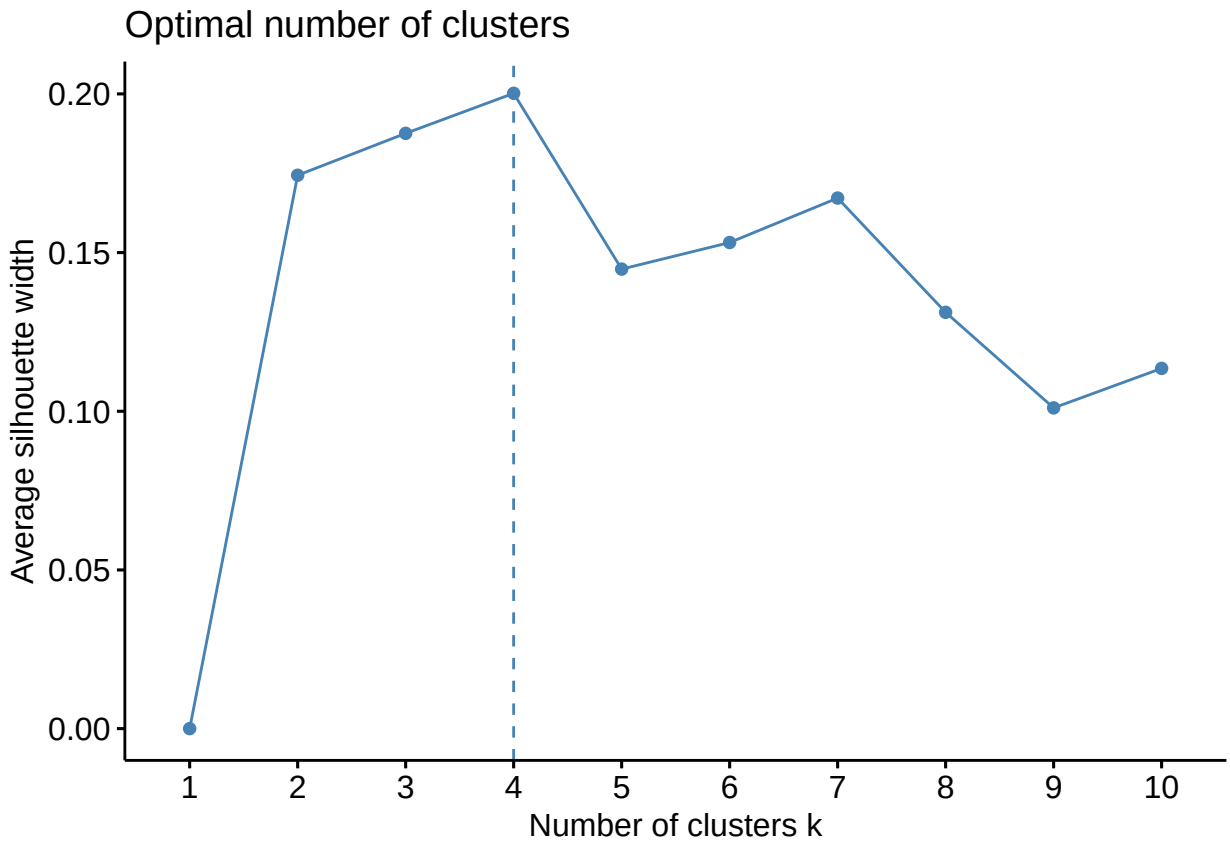
```
##   average    single  complete      ward
## 0.9066583 0.9039645 0.9359926 0.9868039
```

Now looking at divisive hierarchial clustering to compare to the AGNES algorithm in order to see which algorthm is more effective with clustering based of off clustering structure coefficients. Looking below it is evident the divise coefficient for divisive hierarchial clustering is not as high as high as AGNES's comlpete linkage or Ward's linkage. Thus, the best model to use for this admissions data will be the AGNES algorithm using Ward's method. Interesting to note that the divise coefficient appears to be in the middle of the the four different AGNES models.

```
## [1] 0.9270291
```

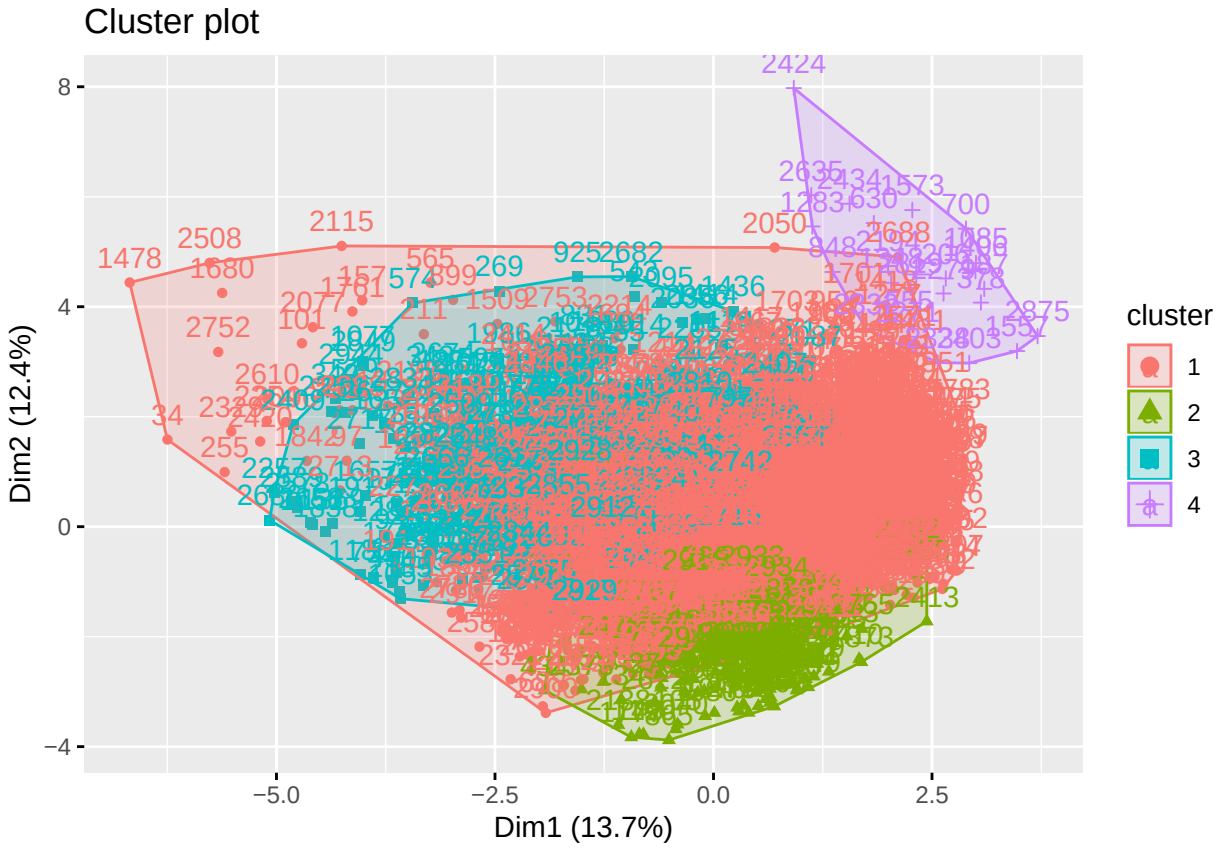
Next, to figure out how to cluster the data I used two methods to find the optimal number of clusters for the data. The first plot below is the elbow method and there does not appear to be a clear optimal cluster. However, it seems that seven clusters seem to show more bend in the "elbow" then any other point. Looking at the model below the elbow method, which is the average silhouette method, we see that the optimal number of clusters is clearly four. Thus, I will primarily use four clusters for my hierarchial clustering but I will also use seven just to compare results and outcomes and see if there are any interesting insights that arise that I could not find with using four clusters.

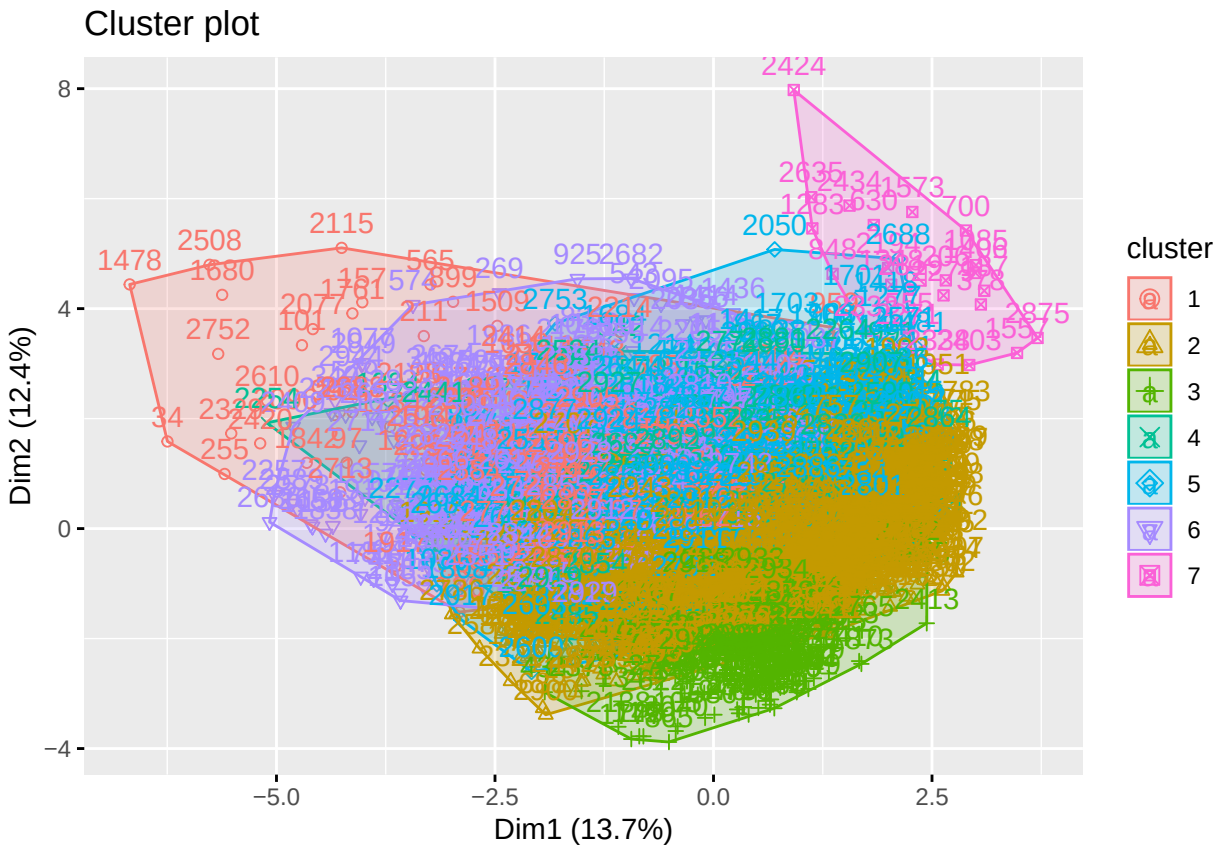




Looking now at cluster plots below, there seems to be alot of ver either in the cluster plot with four clusters

or the cluster plot with seven clusters. This can be an issue because it means that there is not a lot of distinctness between clusters. However, when we look at the descriptive statistics there seems to be some key distinction in AcdmcIndex, ACTComp, GPAREcalc, Decision, EventParticipation and CampusVisits.





I will focus on the descriptive statistics of the hierarchial clustering model that is split up by four clusters due to four clusters being the most clearly observed optimal cluster. Looking at the descritptive statistics of the four clusters seen below, it seems that cluster number four, which contains all of the full ride students, have a mean GPARecalc of 3.91 with a median of 3.93 and that most of those students in this group decided on attending the university based on looking at the mean of Decision which is .67. I think also that it is interesting that group four had an average ACTComp of 32.23 with a standard deviation of 2.3, so very high ACT scores, and that the students in this group on average visited the school with CampusVisits being .54. Furthemore most of the students in group four were on average in the top 5% of their respective schools. Interesting to note is that most of the students by far are grouped in group one. Which on average received a merit award of 19,370 dollars with a standard deviation of 4,710 dollars while having on average a 3.62 GPARecalc and an average Decision of .22 and median of 0, which means more often than not, they decide not to come to the university. Also I noticed with group one that on average these students were a number 3 for AcdmcIndex. Which means that the average accepted university student, which made up the majority of the dataset, did not decide to go to the university. Group one I would describe as the average accepted student, characterized by having around 19,000 merit award with an average GPA of 3.62, plus or minus .33, with an avergae academic index of around 3 and ACTComp of 29.74. Group four I would characterize as the elite group of accepted students who get full ride scholarships, have an average academic index of 1 with an average GPA close to 4.0 and an average ACT composite score of 32 and who have a class rank in the top five percent. Group two I would characterize as the upper end average accepted student with an avergae merit award of 20,000 dollars, with an average academic index of around three and have on average a 3.62, plus or minus .32, GPA with an ACT composite score of a 29 with a standard deviation 1.22. Group three would represent the lowest tier of accpeted students with an average merit award of 18,000 dollars, with the lowest ACT and GPA on average.

```
##
##  Descriptive statistics by group
## group: 1
```

```
##                     vars    n   mean     sd median trimmed  mad   min  max
## PermntCountry*         1 2369   2.00   0.06   2.00    2.00 0.00  1.00    2
## Ethnicity              2 2369   0.21   0.41   0.00    0.14 0.00  0.00    1
## Inquired*              3 2369   1.65   0.48   2.00    1.68 0.00  1.00    2
## DecisionPlan*          4 2369   1.37   0.76   1.00    1.21 0.00  1.00    3
## Legacy                 5 2369   0.10   0.29   0.00    0.00 0.00  0.00    1
## Athlete                6 2369   0.04   0.22   0.00    0.00 0.00  0.00    3
## SportRating*           7 2369   3.00   0.23   3.00    3.00 0.00  1.00    4
## EventParticipation     8 2369   0.33   0.57   0.00    0.23 0.00  0.00    3
## CampusVisits           9 2369   0.42   1.07   0.00    0.13 0.00  0.00    8
## ClassRank             10 2369  33.13  69.12   0.00   15.57 0.00  0.00  655
## ClassSize             11 2369 236.04 308.52   0.00  188.37 0.00  0.00 1569
## GPARecalc             12 2369   3.62   0.33   3.69    3.66 0.37  2.33    4
## ACTComp               13 2369  29.74   2.53  29.00   29.72 1.48 20.00   36
## Decision              14 2369   0.22   0.42   0.00    0.15 0.00  0.00    1
## AcdmcIndex            15 2369   2.63   1.12   3.00    2.60 1.48  1.00    5
## FinancialAid          16 2369   0.69   0.46   1.00    0.73 0.00  0.00    1
## MeritAward            17 2369  19.37   4.71  21.00   20.14 4.45  0.00   28
##                      range   skew kurtosis   se
## PermntCountry*        1.00 -16.12   258.01 0.00
## Ethnicity             1.00   1.43     0.06 0.01
## Inquired*             1.00  -0.62    -1.62 0.01
## DecisionPlan*         2.00   1.63     0.73 0.02
## Legacy                1.00   2.74     5.53 0.01
## Athlete               3.00   7.39    70.36 0.00
## SportRating*          3.00  -3.12    43.94 0.00
## EventParticipation    3.00   1.66     2.39 0.01
## CampusVisits          8.00   2.65     6.76 0.02
## ClassRank           655.00   3.27    13.95 1.42
## ClassSize          1569.00   1.16     0.80 6.34
## GPARecalc             1.67  -0.67    -0.42 0.01
## ACTComp              16.00   0.04     0.42 0.05
## Decision              1.00   1.34    -0.20 0.01
## AcdmcIndex            4.00   0.15    -0.78 0.02
## FinancialAid          1.00  -0.81    -1.35 0.01
## MeritAward           28.00  -2.28     6.47 0.10
## ------------------------------------------------------------
## group: 2
##                     vars   n  mean     sd median trimmed  mad   min max  range
## PermntCountry*         1 180  1.00   0.00   1.00    1.00 0.00  1.00   1   0.00
## Ethnicity              2 180  0.36   0.48   0.00    0.32 0.00  0.00   1   1.00
## Inquired*              3 180  1.39   0.49   1.00    1.37 0.00  1.00   2   1.00
## DecisionPlan*          4 180  1.81   0.98   1.00    1.76 0.00  1.00   3   2.00
## Legacy                 5 180  0.05   0.22   0.00    0.00 0.00  0.00   1   1.00
## Athlete                6 180  0.05   0.26   0.00    0.00 0.00  0.00   2   2.00
## SportRating*           7 180  2.98   0.24   3.00    3.00 0.00  1.00   4   3.00
## EventParticipation     8 180  0.04   0.19   0.00    0.00 0.00  0.00   1   1.00
## CampusVisits           9 180  0.11   0.53   0.00    0.00 0.00  0.00   3   3.00
## ClassRank             10 180  1.78   9.31   0.00    0.04 0.00  0.00  96  96.00
## ClassSize             11 180 32.18 129.74   0.00    1.74 0.00  0.00 883 883.00
## GPARecalc             12 180  3.62   0.32   3.67    3.65 0.35  2.61   4   1.39
## ACTComp               13 180 29.02   1.44  29.00   28.98 0.00 24.00  35  11.00
## Decision              14 180  0.15   0.36   0.00    0.06 0.00  0.00   1   1.00
## AcdmcIndex            15 180  2.94   0.52   3.00    3.00 0.00  1.00   5   4.00
```

```
## FinancialAid         16 180  0.58   0.49   1.00    0.60 0.00  0.00   1   1.00
## MeritAward           17 180 20.37   5.79  21.00   20.53 4.45  0.00  38  38.00
##                     skew kurtosis   se
## PermntCountry*       NaN      NaN 0.00
## Ethnicity           0.60    -1.65 0.04
## Inquired*           0.43    -1.83 0.04
## DecisionPlan*       0.39    -1.85 0.07
## Legacy              4.10    14.85 0.02
## Athlete             5.71    34.29 0.02
## SportRating*       -3.98    35.31 0.02
## EventParticipation  4.73    20.49 0.01
## CampusVisits        4.67    20.67 0.04
## ClassRank           7.63    65.00 0.69
## ClassSize           5.12    27.37 9.67
## GPARecalc          -0.75    -0.12 0.02
## ACTComp             0.84     5.82 0.11
## Decision            1.94     1.79 0.03
## AcdmcIndex         -0.80     6.68 0.04
## FinancialAid       -0.34    -1.90 0.04
## MeritAward         -0.42     2.38 0.43
## ------------------------------------------------------------
## group: 3
##                     vars   n   mean     sd median trimmed  mad   min max  range
## PermntCountry*         1 373   2.00   0.00    2.0    2.00 0.00  2.00   2   0.00
## Ethnicity              2 373   0.17   0.38    0.0    0.09 0.00  0.00   1   1.00
## Inquired*              3 373   1.62   0.49    2.0    1.65 0.00  1.00   2   1.00
## DecisionPlan*          4 373   1.27   0.66    1.0    1.09 0.00  1.00   3   2.00
## Legacy                 5 373   0.08   0.27    0.0    0.00 0.00  0.00   1   1.00
## Athlete                6 373   1.50   0.70    1.0    1.38 0.00  1.00   3   2.00
## SportRating*           7 373   2.34   1.35    2.0    2.30 1.48  1.00   4   3.00
## EventParticipation     8 373   0.18   0.41    0.0    0.09 0.00  0.00   2   2.00
## CampusVisits           9 373   0.71   1.21    0.0    0.47 0.00  0.00   5   5.00
## ClassRank             10 373  23.73  41.77    0.0   13.74 0.00  0.00 230 230.00
## ClassSize             11 373 209.59 269.58    0.0  170.63 0.00  0.00 930 930.00
## GPARecalc             12 373   3.61   0.35    3.7    3.65 0.36  2.53   4   1.47
## ACTComp               13 373  28.67   2.70   29.0   28.68 2.97 17.00  35  18.00
## Decision              14 373   0.24   0.43    0.0    0.18 0.00  0.00   1   1.00
## AcdmcIndex            15 373   2.85   1.10    3.0    2.85 1.48  1.00   5   4.00
## FinancialAid          16 373   0.70   0.46    1.0    0.75 0.00  0.00   1   1.00
## MeritAward            17 373  18.01   5.31   21.0   18.83 4.45  0.00  24  24.00
##                     skew kurtosis    se
## PermntCountry*       NaN      NaN  0.00
## Ethnicity           1.71     0.93  0.02
## Inquired*          -0.49    -1.77  0.03
## DecisionPlan*       2.12     2.68  0.03
## Legacy              3.07     7.46  0.01
## Athlete             1.04    -0.26  0.04
## SportRating*        0.29    -1.72  0.07
## EventParticipation  2.08     3.50  0.02
## CampusVisits        1.44     0.95  0.06
## ClassRank           2.12     4.27  2.16
## ClassSize           0.84    -0.76 13.96
## GPARecalc          -0.81    -0.23  0.02
## ACTComp            -0.23     0.69  0.14
```

```
## Decision            1.20    -0.55  0.02
## AcdmcIndex          0.09    -0.71  0.06
## FinancialAid       -0.85    -1.27  0.02
## MeritAward         -1.87     3.94  0.27
## ------------------------------------------------------------
## group: 4
##                     vars  n     mean     sd   median  trimmed    mad      min
## PermntCountry*         1 24     2.00   0.00     2.00     2.00   0.00     2.00
## Ethnicity              2 24     0.12   0.34     0.00     0.05   0.00     0.00
## Inquired*              3 24     1.79   0.41     2.00     1.85   0.00     1.00
## DecisionPlan*          4 24     1.04   0.20     1.00     1.00   0.00     1.00
## Legacy                 5 24     0.21   0.41     0.00     0.15   0.00     0.00
## Athlete                6 24     0.42   0.93     0.00     0.20   0.00     0.00
## SportRating*           7 24     3.08   0.58     3.00     3.10   0.00     1.00
## EventParticipation     8 24     1.42   0.65     1.00     1.30   0.00     1.00
## CampusVisits           9 24     0.54   0.93     0.00     0.40   0.00     0.00
## ClassRank             10 24    15.58  22.52     3.00    11.35   4.45     0.00
## ClassSize             11 24   335.00 381.55   158.00   289.85 234.25     0.00
## GPARecalc             12 24     3.91   0.10     3.93     3.92   0.10     3.69
## ACTComp               13 24    32.21   2.32    33.00    32.25   2.22    29.00
## Decision              14 24     0.67   0.48     1.00     0.70   0.00     0.00
## AcdmcIndex            15 24     1.25   0.44     1.00     1.20   0.00     1.00
## FinancialAid          16 24     0.79   0.41     1.00     0.85   0.00     0.00
## MeritAward            17 24 57000.00   0.00 57000.00 57000.00   0.00 57000.00
##                      max   range  skew kurtosis    se
## PermntCountry*         2    0.00   NaN      NaN  0.00
## Ethnicity              1    1.00  2.13     2.64  0.07
## Inquired*              2    1.00 -1.35    -0.19  0.08
## DecisionPlan*          2    1.00  4.30    17.24  0.04
## Legacy                 1    1.00  1.35    -0.19  0.08
## Athlete                3    3.00  1.95     2.39  0.19
## SportRating*           4    3.00 -1.25     4.78  0.12
## EventParticipation     3    2.00  1.19     0.13  0.13
## CampusVisits           3    3.00  1.28     0.11  0.19
## ClassRank             87   87.00  1.58     2.04  4.60
## ClassSize           1449 1449.00  1.01     0.50 77.88
## GPARecalc              4    0.31 -0.89    -0.62  0.02
## ACTComp               35    6.00 -0.34    -1.45  0.47
## Decision               1    1.00 -0.66    -1.62  0.10
## AcdmcIndex             2    1.00  1.08    -0.86  0.09
## FinancialAid           1    1.00 -1.35    -0.19  0.08
## MeritAward         57000    0.00   NaN      NaN  0.00
```

K-means Clustering

For k-means clustering I began with four clusters. I decided on the four clusters based off of the thought process that a two clustered model would be too simple and that based off of the elbow method in hierarchial clustering models, seven clusters seemed to be the maximum number of optimal clusters without reducing too much sensitivity if I were to perform the clustering on a test set. I varied the number of starts and number of iterations to find when the observations would be clustered the consistently in the same cluster and found that performing 100 starts on a maximum of 100 iterations was the point in which the observations had begun to be consistently clustered in the same cluster. The descriptive statistics for this trial run of choosing K based of of theory are given below. Is is interesting just at first glance to compare how group one

for hierarchial clustering with four clusters compared to group four with this initial k-means model in that majority group lost about 400 members from with k-means clustering.

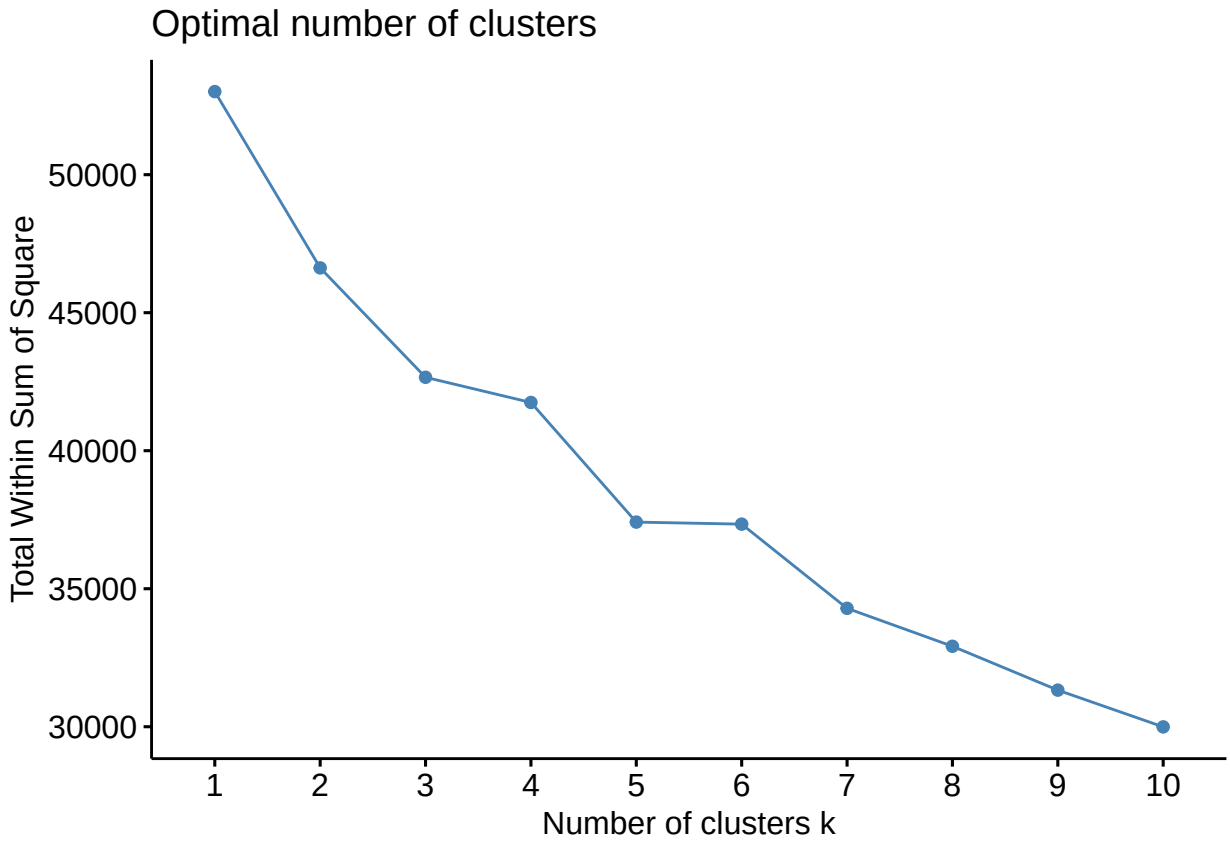
```
##
##  Descriptive statistics by group
## group: 1
##                     vars  n     mean     sd   median  trimmed    mad      min
## PermntCountry*         1 24     2.00   0.00     2.00     2.00   0.00     2.00
## Ethnicity              2 24     0.12   0.34     0.00     0.05   0.00     0.00
## Inquired*              3 24     1.79   0.41     2.00     1.85   0.00     1.00
## DecisionPlan*          4 24     1.04   0.20     1.00     1.00   0.00     1.00
## Legacy                 5 24     0.21   0.41     0.00     0.15   0.00     0.00
## Athlete                6 24     0.42   0.93     0.00     0.20   0.00     0.00
## SportRating*           7 24     3.08   0.58     3.00     3.10   0.00     1.00
## EventParticipation     8 24     1.42   0.65     1.00     1.30   0.00     1.00
## CampusVisits           9 24     0.54   0.93     0.00     0.40   0.00     0.00
## ClassRank             10 24    15.58  22.52     3.00    11.35   4.45     0.00
## ClassSize             11 24   335.00 381.55   158.00   289.85 234.25     0.00
## GPARecalc             12 24     3.91   0.10     3.93     3.92   0.10     3.69
## ACTComp               13 24    32.21   2.32    33.00    32.25   2.22    29.00
## Decision              14 24     0.67   0.48     1.00     0.70   0.00     0.00
## AcdmcIndex            15 24     1.25   0.44     1.00     1.20   0.00     1.00
## FinancialAid          16 24     0.79   0.41     1.00     0.85   0.00     0.00
## MeritAward            17 24 57000.00   0.00 57000.00 57000.00   0.00 57000.00
## mdl4$cluster          18 24     1.00   0.00     1.00     1.00   0.00     1.00
##                       max   range  skew kurtosis    se
## PermntCountry*          2    0.00   NaN      NaN  0.00
## Ethnicity               1    1.00  2.13     2.64  0.07
## Inquired*               2    1.00 -1.35    -0.19  0.08
## DecisionPlan*           2    1.00  4.30    17.24  0.04
## Legacy                  1    1.00  1.35    -0.19  0.08
## Athlete                 3    3.00  1.95     2.39  0.19
## SportRating*            4    3.00 -1.25     4.78  0.12
## EventParticipation      3    2.00  1.19     0.13  0.13
## CampusVisits            3    3.00  1.28     0.11  0.19
## ClassRank              87   87.00  1.58     2.04  4.60
## ClassSize            1449 1449.00  1.01     0.50 77.88
## GPARecalc               4    0.31 -0.89    -0.62  0.02
## ACTComp                35    6.00 -0.34    -1.45  0.47
## Decision                1    1.00 -0.66    -1.62  0.10
## AcdmcIndex              2    1.00  1.08    -0.86  0.09
## FinancialAid            1    1.00 -1.35    -0.19  0.08
## MeritAward          57000    0.00   NaN      NaN  0.00
## mdl4$cluster            1    0.00   NaN      NaN  0.00
## ------------------------------------------------------------
## group: 2
##                     vars    n  mean    sd median trimmed  mad   min max  range
## PermntCountry*         1 1882  1.90  0.29   2.00    2.00 0.00  1.00   2   1.00
## Ethnicity              2 1882  0.22  0.42   0.00    0.15 0.00  0.00   1   1.00
## Inquired*              3 1882  1.63  0.48   2.00    1.66 0.00  1.00   2   1.00
## DecisionPlan*          4 1882  1.38  0.77   1.00    1.23 0.00  1.00   3   2.00
## Legacy                 5 1882  0.09  0.29   0.00    0.00 0.00  0.00   1   1.00
## Athlete                6 1882  0.22  0.59   0.00    0.06 0.00  0.00   3   3.00
## SportRating*           7 1882  2.94  0.54   3.00    3.00 0.00  1.00   4   3.00
```

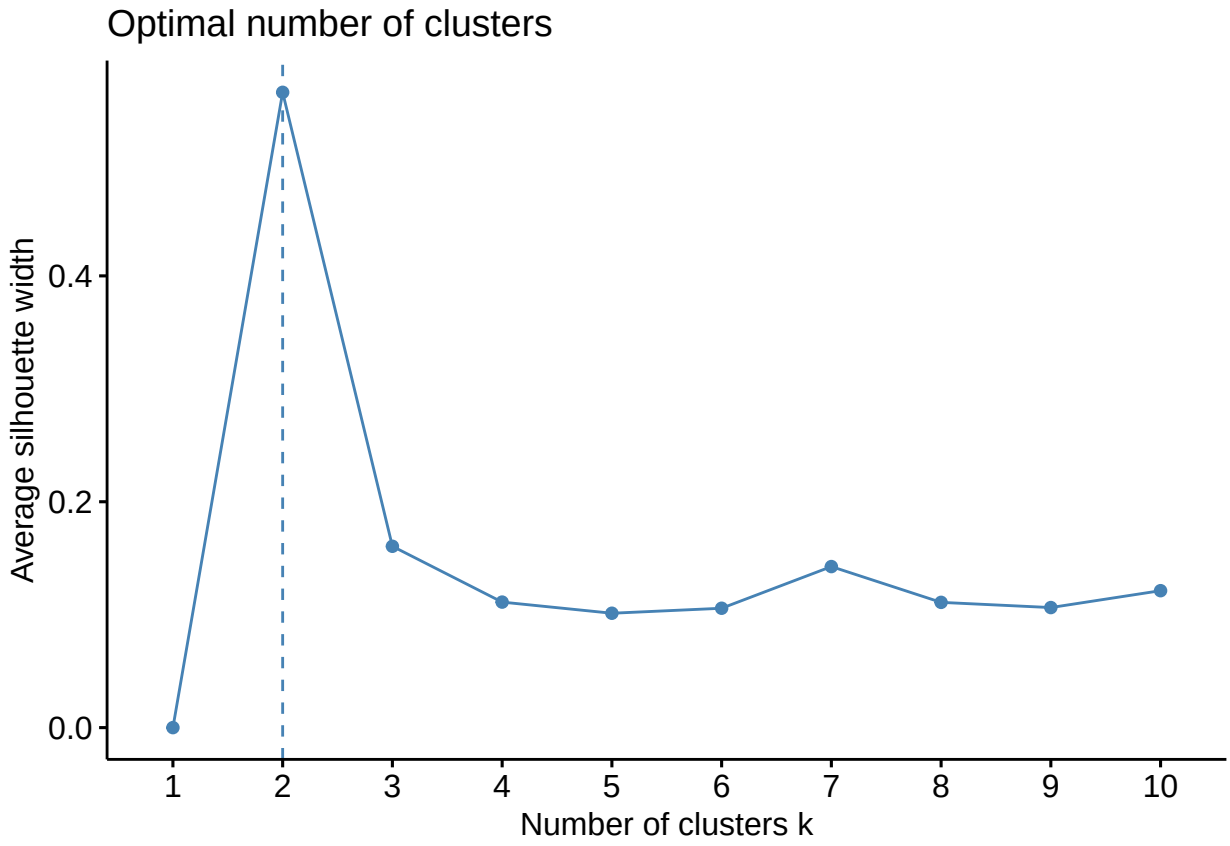
```
## EventParticipation    8 1882  0.26  0.52   0.00    0.16 0.00  0.00   3   3.00
## CampusVisits          9 1882  0.43  1.09   0.00    0.13 0.00  0.00   8   8.00
## ClassRank            10 1882  2.64 10.59   0.00    0.16 0.00  0.00 131 131.00
## ClassSize            11 1882 19.82 52.57   0.00    4.20 0.00  0.00 263 263.00
## GPARecalc            12 1882  3.59  0.34   3.64    3.62 0.40  2.33   4   1.67
## ACTComp              13 1882 29.50  2.46  29.00   29.49 1.48 21.00  36  15.00
## Decision             14 1882  0.20  0.40   0.00    0.12 0.00  0.00   1   1.00
## AcdmcIndex           15 1882  2.78  1.08   3.00    2.77 1.48  1.00   5   4.00
## FinancialAid         16 1882  0.64  0.48   1.00    0.68 0.00  0.00   1   1.00
## MeritAward           17 1882 19.08  4.99  21.00   19.86 4.45  0.00  38  38.00
## mdl4$cluster         18 1882  2.00  0.00   2.00    2.00 0.00  2.00   2   0.00
##                      skew kurtosis   se
## PermntCountry*     -2.74     5.50 0.01
## Ethnicity           1.33    -0.24 0.01
## Inquired*          -0.52    -1.73 0.01
## DecisionPlan*       1.57     0.51 0.02
## Legacy              2.79     5.79 0.01
## Athlete             3.01     9.00 0.01
## SportRating*       -1.93     6.38 0.01
## EventParticipation  1.96     3.59 0.01
## CampusVisits        2.61     6.56 0.03
## ClassRank           6.04    43.89 0.24
## ClassSize           2.79     7.00 1.21
## GPARecalc          -0.60    -0.52 0.01
## ACTComp             0.10     0.49 0.06
## Decision            1.51     0.29 0.01
## AcdmcIndex          0.04    -0.60 0.02
## FinancialAid       -0.60    -1.64 0.01
## MeritAward         -1.83     5.08 0.11
## mdl4$cluster         NaN      NaN 0.00
## ------------------------------------------------------------
## group: 3
##                     vars   n   mean     sd median trimmed    mad    min  max
## PermntCountry*         1 218   1.99   0.12   2.00    2.00   0.00   1.00    2
## Ethnicity              2 218   0.19   0.39   0.00    0.11   0.00   0.00    1
## Inquired*              3 218   1.62   0.49   2.00    1.65   0.00   1.00    2
## DecisionPlan*          4 218   1.34   0.74   1.00    1.18   0.00   1.00    3
## Legacy                 5 218   0.11   0.31   0.00    0.01   0.00   0.00    1
## Athlete                6 218   0.25   0.61   0.00    0.10   0.00   0.00    3
## SportRating*           7 218   2.82   0.67   3.00    2.95   0.00   1.00    4
## EventParticipation     8 218   0.39   0.66   0.00    0.27   0.00   0.00    3
## CampusVisits           9 218   0.35   0.92   0.00    0.11   0.00   0.00    5
## ClassRank             10 218 142.98 127.66 112.00  126.64 114.16   1.00  655
## ClassSize             11 218 882.91 210.21 795.00  843.44  88.96 655.00 1569
## GPARecalc             12 218   3.61   0.32   3.65    3.63   0.40   2.71    4
## ACTComp               13 218  29.44   2.42  29.00   29.50   1.48  21.00   35
## Decision              14 218   0.24   0.43   0.00    0.18   0.00   0.00    1
## AcdmcIndex            15 218   2.71   1.14   3.00    2.70   1.48   1.00    5
## FinancialAid          16 218   0.77   0.42   1.00    0.84   0.00   0.00    1
## MeritAward            17 218  19.33   4.60  21.00   20.02   4.45   0.00   27
## mdl4$cluster          18 218   3.00   0.00   3.00    3.00   0.00   3.00    3
##                     range  skew kurtosis    se
## PermntCountry*       1.00 -8.29    67.03  0.01
## Ethnicity            1.00  1.59     0.52  0.03
```

```
## Inquired*            1.00 -0.51    -1.75  0.03
## DecisionPlan*        2.00  1.75     1.12  0.05
## Legacy               1.00  2.55     4.53  0.02
## Athlete              3.00  2.69     7.23  0.04
## SportRating*         3.00 -1.77     2.98  0.05
## EventParticipation   3.00  1.69     2.58  0.04
## CampusVisits         5.00  2.71     6.94  0.06
## ClassRank          654.00  1.21     1.47  8.65
## ClassSize          914.00  1.61     1.71 14.24
## GPARecalc            1.29 -0.55    -0.72  0.02
## ACTComp             14.00 -0.27     0.91  0.16
## Decision             1.00  1.19    -0.59  0.03
## AcdmcIndex           4.00 -0.02    -0.88  0.08
## FinancialAid         1.00 -1.28    -0.37  0.03
## MeritAward          27.00 -1.97     5.43  0.31
## mdl4$cluster         0.00   NaN      NaN  0.00
## ------------------------------------------------------------
## group: 4
##                     vars   n   mean     sd median trimmed    mad    min max
## PermntCountry*         1 822   1.99   0.08   2.00    2.00   0.00   1.00   2
## Ethnicity              2 822   0.20   0.40   0.00    0.12   0.00   0.00   1
## Inquired*              3 822   1.64   0.48   2.00    1.67   0.00   1.00   2
## DecisionPlan*          4 822   1.39   0.78   1.00    1.24   0.00   1.00   3
## Legacy                 5 822   0.08   0.27   0.00    0.00   0.00   0.00   1
## Athlete                6 822   0.23   0.56   0.00    0.09   0.00   0.00   3
## SportRating*           7 822   2.89   0.60   3.00    2.99   0.00   1.00   4
## EventParticipation     8 822   0.34   0.56   0.00    0.25   0.00   0.00   3
## CampusVisits           9 822   0.47   1.06   0.00    0.20   0.00   0.00   5
## ClassRank             10 822  62.66  63.80  43.00   52.06  48.18   1.00 355
## ClassSize             11 822 502.89 113.90 513.00  506.50 139.36 261.00 708
## GPARecalc             12 822   3.70   0.29   3.79    3.74   0.30   2.52   4
## ACTComp               13 822  29.72   2.71  29.00   29.73   1.48  17.00  36
## Decision              14 822   0.26   0.44   0.00    0.20   0.00   0.00   1
## AcdmcIndex            15 822   2.44   1.06   2.00    2.39   1.48   1.00   5
## FinancialAid          16 822   0.75   0.44   1.00    0.81   0.00   0.00   1
## MeritAward            17 822  19.65   4.70  21.00   20.40   4.45   0.00  28
## mdl4$cluster          18 822   4.00   0.00   4.00    4.00   0.00   4.00   4
##                     range   skew kurtosis   se
## PermntCountry*       1.00 -12.68   159.01 0.00
## Ethnicity            1.00   1.53     0.34 0.01
## Inquired*            1.00  -0.57    -1.68 0.02
## DecisionPlan*        2.00   1.53     0.42 0.03
## Legacy               1.00   3.05     7.33 0.01
## Athlete              3.00   2.75     7.87 0.02
## SportRating*         3.00  -1.79     4.56 0.02
## EventParticipation   3.00   1.49     1.67 0.02
## CampusVisits         5.00   2.20     4.00 0.04
## ClassRank          354.00   1.55     2.47 2.23
## ClassSize          447.00  -0.24    -1.01 3.97
## GPARecalc            1.48  -0.92     0.16 0.01
## ACTComp             19.00  -0.12     0.77 0.09
## Decision             1.00   1.08    -0.83 0.02
## AcdmcIndex           4.00   0.28    -0.73 0.04
## FinancialAid         1.00  -1.13    -0.71 0.02
```

```
## MeritAward          28.00  -2.46     7.60 0.16
## mdl4$cluster         0.00    NaN      NaN 0.00
```

However, when I performed the elbow method and average silhouette for k-means clustering I found that either two or five clusters is the optimal amount of clusters for this data. The results of the elbow method and silhouette method can be seen below. The elbow method has a sgingficant elbow at both when K is five and when K is eight. Eight clusters seems to high, especially with keeping in mind with the optimal k of four for the hierarchical clustering model, so that is why five clusters seems to be the most appropriate maximum amount of clusters.





I decided to go with a k-means model split up into two clusters due to the observation from the hierarchical clustering performed earlier that other than the top group of students, the rest of the toher students were actaully very similar. This gives me confidence that a two cluster k-means model is the most appropriate way of splitting the data. Looking at the model below it is evident that group one reperesnets the top gorup of accepted students with a median of 21,000 dollars of merit award with a median GPA of 3.70 and average of 3.63, with an average ACT composite score of 29.75 and median of 29.00, with an average academic index 2.61 and median of 3.00. Interesting to note is that approximately 85% of the students belonged in group one. Group two represents the lower end of the tier with probably with a median merit award of 17,000 dollars, an average ACT composite score of 28.65, an average GPA of 3.57.

```
##
##  Descriptive statistics by group
## group: 1
##    vars    n  mean    sd median trimmed  mad   min max  range   skew kurtosis
## 1     1 1884  1.90  0.30   2.00    2.00 0.00  1.00   2   1.00  -2.72     5.39
## 2     2 1884  0.22  0.42   0.00    0.15 0.00  0.00   1   1.00   1.33    -0.24
## 3     3 1884  1.62  0.48   2.00    1.66 0.00  1.00   2   1.00  -0.51    -1.74
## 4     4 1884  1.38  0.78   1.00    1.23 0.00  1.00   3   2.00   1.56     0.49
## 5     5 1884  0.09  0.29   0.00    0.00 0.00  0.00   1   1.00   2.79     5.80
## 6     6 1884  0.22  0.59   0.00    0.06 0.00  0.00   3   3.00   3.01     9.02
## 7     7 1884  2.94  0.54   3.00    3.00 0.00  1.00   4   3.00  -1.93     6.39
## 8     8 1884  0.26  0.52   0.00    0.16 0.00  0.00   3   3.00   1.96     3.60
## 9     9 1884  0.43  1.09   0.00    0.13 0.00  0.00   8   8.00   2.61     6.57
## 10   10 1884  2.64 10.58   0.00    0.16 0.00  0.00 131 131.00   6.04    43.94
## 11   11 1884 20.17 53.66   0.00    4.32 0.00  0.00 358 358.00   2.86     7.73
## 12   12 1884  3.59  0.34   3.64    3.62 0.40  2.33   4   1.67  -0.60    -0.52
## 13   13 1884 29.50  2.46  29.00   29.49 1.48 21.00  36  15.00   0.10     0.49
```

```
## 14   14 1884  0.20  0.40   0.00    0.12 0.00  0.00   1   1.00   1.51     0.29
## 15   15 1884  2.78  1.08   3.00    2.78 1.48  1.00   5   4.00   0.04    -0.60
## 16   16 1884  0.64  0.48   1.00    0.68 0.00  0.00   1   1.00  -0.60    -1.64
## 17   17 1884 19.08  4.99  21.00   19.86 4.45  0.00  38  38.00  -1.83     5.07
## 18   18 1884  2.00  0.07   2.00    2.00 0.00  2.00   4   2.00  30.62   936.00
## 19   19 1884  2.00  0.03   2.00    2.00 0.00  2.00   3   1.00  30.62   936.00
## 20   20 1884  4.00  0.03   4.00    4.00 0.00  3.00   4   1.00 -30.62   936.00
## 21   21 1884  4.00  0.07   4.00    4.00 0.00  2.00   4   2.00 -30.62   936.00
## 22   22 1884  4.00  0.07   4.00    4.00 0.00  2.00   4   2.00 -30.62   936.00
## 23   23 1884  1.00  0.00   1.00    1.00 0.00  1.00   1   0.00    NaN      NaN
##       se
## 1   0.01
## 2   0.01
## 3   0.01
## 4   0.02
## 5   0.01
## 6   0.01
## 7   0.01
## 8   0.01
## 9   0.03
## 10 0.24
## 11 1.24
## 12 0.01
## 13 0.06
## 14 0.01
## 15 0.02
## 16 0.01
## 17 0.11
## 18 0.00
## 19 0.00
## 20 0.00
## 21 0.00
## 22 0.00
## 23 0.00
## ------------------------------------------------------------
## group: 2
##    vars    n    mean       sd median trimmed    mad   min   max    range   skew
## 1     1 1062    1.99    0.07   2.00    2.00   0.00  1.00     2     1.00 -13.17
## 2     2 1062    0.19    0.39   0.00    0.12   0.00  0.00     1     1.00   1.55
## 3     3 1062    1.64    0.48   2.00    1.67   0.00  1.00     2     1.00  -0.58
## 4     4 1062    1.37    0.76   1.00    1.21   0.00  1.00     3     2.00   1.62
## 5     5 1062    0.09    0.29   0.00    0.00   0.00  0.00     1     1.00   2.87
## 6     6 1062    0.24    0.58   0.00    0.09   0.00  0.00     3     3.00   2.75
## 7     7 1062    2.88    0.61   3.00    2.99   0.00  1.00     4     3.00  -1.79
## 8     8 1062    0.38    0.61   0.00    0.27   0.00  0.00     3     3.00   1.54
## 9     9 1062    0.45    1.03   0.00    0.19   0.00  0.00     5     5.00   2.27
## 10   10 1062   78.20   87.30  49.00   62.35  57.82  0.00   655   655.00   2.07
## 11   11 1062  577.39  216.67 559.00  557.97 169.02  0.00  1569  1569.00   1.28
## 12   12 1062    3.69    0.30   3.77    3.72   0.30  2.52     4     1.48  -0.87
## 13   13 1062   29.72    2.67  29.00   29.74   1.48 17.00    36    19.00  -0.13
## 14   14 1062    0.27    0.44   0.00    0.21   0.00  0.00     1     1.00   1.05
## 15   15 1062    2.46    1.09   2.00    2.42   1.48  1.00     5     4.00   0.25
## 16   16 1062    0.75    0.43   1.00    0.82   0.00  0.00     1     1.00  -1.17
## 17   17 1062 1307.26 8472.47  21.00   20.42   4.45  0.00 57000 57000.00   6.42
```

```
## 18   18 1062    3.73    0.58   4.00    3.84   0.00  1.00     4     3.00  -2.70
## 19   19 1062    3.16    0.52   3.00    3.13   0.00  1.00     4     3.00  -0.78
## 20   20 1062    2.75    0.48   3.00    2.84   0.00  1.00     3     2.00  -1.73
## 21   21 1062    1.82    0.44   2.00    1.87   0.00  1.00     3     2.00  -0.81
## 22   22 1062    1.82    0.44   2.00    1.87   0.00  1.00     3     2.00  -0.81
## 23   23 1062    2.00    0.00   2.00    2.00   0.00  2.00     2     0.00    NaN
##    kurtosis     se
## 1    171.68   0.00
## 2      0.41   0.01
## 3     -1.67   0.01
## 4      0.70   0.02
## 5      6.26   0.01
## 6      7.75   0.02
## 7      4.23   0.02
## 8      2.06   0.02
## 9      4.41   0.03
## 10     5.98   2.68
## 11     3.72   6.65
## 12    -0.03   0.01
## 13     0.74   0.08
## 14    -0.90   0.01
## 15    -0.79   0.03
## 16    -0.64   0.01
## 17    39.19 259.98
## 18     8.73   0.02
## 19     5.14   0.02
## 20     2.14   0.01
## 21     0.61   0.01
## 22     0.61   0.01
## 23      NaN   0.00
```

Conclusion

This project brought about some helpful insights into the admissions data. Learend from the hierarchical clustering there are four distinct groups in the data. That the top tier group, group four, were quite distinct from the other three groups. While the other three remaining groups, although still able to be grouped and ordered in a logical way, were very similar. Thus giving the notion that there are two bigger overall groups in the data, the top tiered students and the remainig students who were all very similar. However, with the distinct groups you could assign new students to a group based off of their GPA and other factors and see the likelihood of them going to the university through looking at the mean and median of the variable Decision for that particular group. Furthermore, with the k-means clustering analysis it was evident that there are in fact two distinct groups within the admissions data with distinclty different means and medians for merit awards, GPA, and other categories. Based on the results and insights of the hierarchical and k-means clustering models, it is evident that it is best to use both models to get the full picture of the data. The hierarchical clustering providing the more detailed understanding of the data while the k-means cluster provided the macro view of the data.

Self Reflection

As I reflect on this assignment one problem that I noticed that began to give me a little trouble was having the value of 57000 in the MeritAward variable. I should have changed it to 57 in order to make the means and medinas more readable for the descriptive statistics of each group. Furthermore, I believe that I could

have derived more helpful insights if I were looking at the descriptive statistics of more variables. My choice to roughly focus on MeritAward, AcdmcIndex, ACTComp, GPAREcalc, Decision, EventParticipation and CampusVisits, although helpful in providing guidance so that I do not waste time with getting lost in descriptive statistics, could have caused me to miss some valuable and important insights. My cleaning process of the data I thought was appropriate and was able to retain the variables that I could retain and removed variables that I thought were not able to convert into numeric form without misrepresenting the data. Overall I believe that I was able to glean some interesting insights into the admissions data.