David Hudson

Stat 463-001

Inter-Major Marriage Rates

In my search for a data set to exam I told myself I wanted to choose a data set that no one else would want to choose. I also wanted to choose a data set that would have the rest of the class interested in what I was talking about whether it be truly scientific or just fun facts. I was searching through different discussion board when I came across my data set; the title indicated a table comparing the majors of college educated women and the major of the men they married. This intrigued me as it seemed different than the data sets usually used for projects and would allow me to possibly use clustering, a method I have been wanting to apply to data.

The initial data set came as a matrix of numbers with the title of majors lining the axes; the values within the matrix were scores rating their marriage rates relative to the other values in the matrix. This was chosen as opposed to just regular marriage values as some majors had far greater populations than others. The score was calculated as such “=(Amount of marriages of the x and y majors)/((Total Males Married/Total For All)*Total Females)” , we can just call this the “marriage score” for the continuation of this paper. The data set also came with a list of values corresponding to the rate in which each gender of each major married a non-degree holding student, we can called that the “married-down rate”. To provide even more categorical data I looked up what school each major would be a part of if it were offered at mason, or what it currently is at mason, and labeled this as its category. I also added whether the major is a Bachelor of Arts or a Bachelor of Science. With this information I was able to compare and contrast

the different groups and individual majors in order to make a more complete analysis of my data.

Below in **figure 1** is a screen shot of just a portion of the data I used for better understanding. For each major there are it's categorical variables, the population of males, the population of females, the percent of both genders married down, and the rates of marriage for each other major.

Major	Degree	Category	Men	MenMarDown	Women	WomMarDown	Agriculture	Environment and Natur	Architecture	Communications	Computer and Informat	Education Administrat	Engineering
Agri	BS	Engineering	43929.00	0.30	40446.00	0.43	19.51651	3.504466	0.498501	0.357699	1.014008	0.646596	1.070559
Enviro	BS	Science	37876.00	0.18	24904.00	0.30	1.864173	11.55491	2.383426	0.874357	0.584875	0.44157	0.869117
Architect	BS	Art	34380.00	0.17	21979.00	0.25	0.220773	0	18.18729	0.43148	1.155189	0.321757	1.635545
Comm	BA	Human	145729.00	0.25	251921.00	0.31	0.727295	0.939003	0.431051	2.638238	0.563014	0.797664	0.743946
CompSci	BS	Engineering	230407.00	0.21	81368.00	0.25	0.350781	0.215761	0.394385	0.809741	4.710771	0.111637	2.10655
EducAdmin	Either	Other	148299.00	0.27	522001.00	0.42	1.734364	0.941789	0.726209	0.948306	0.763166	3.280109	0.737842
ENGINEERING	BS	Engineering	492375.00	0.18	129090.00	0.19	0.165309	0.583375	0.871457	0.411006	1.763247	0.304223	3.502022
EnginTech	BS	Engineering	38029.00	0.25	13237.00	0.18	1.28325	0	0	0.137916	2.239946	1.247404	2.499208
LingForeign	BA	Human	20053.00	0.22	62664.00	0.28	0.233519	1.632425	1.179413	0.859272	0.881853	0.751482	0.922621
English	BA	Human	75059.00	0.20	167581.00	0.29	0.867294	1.167329	0.765922	1.354151	1.120636	0.656662	0.71057

Figure 1: Screen Shot of Data

So with my data I thought the first thing I should do is simply put the data on a scatter plot and see what kind of trends I start to see. Using ggplot I plotted each major by their rate of men marrying down and their rate of women marrying down. I then coordinated the color of each dot to that major's category, this can be shown in **figure 2**. From this graph we can begin to analyze the behavior of each major. The first thing to notice is that while the plots seem to form a linear relationship with about a slope of 1, if we look at the scales for the two axes we notice that the women scale is larger than the men scale. This is information that can prove to be useful for late analysis. We also see that the engineering majors and science majors are mostly populating the lower values

on both axes. While the humanities have the most amount of points, they also have a much greater distribution with an outlier reaching the upper limits of the scales on both

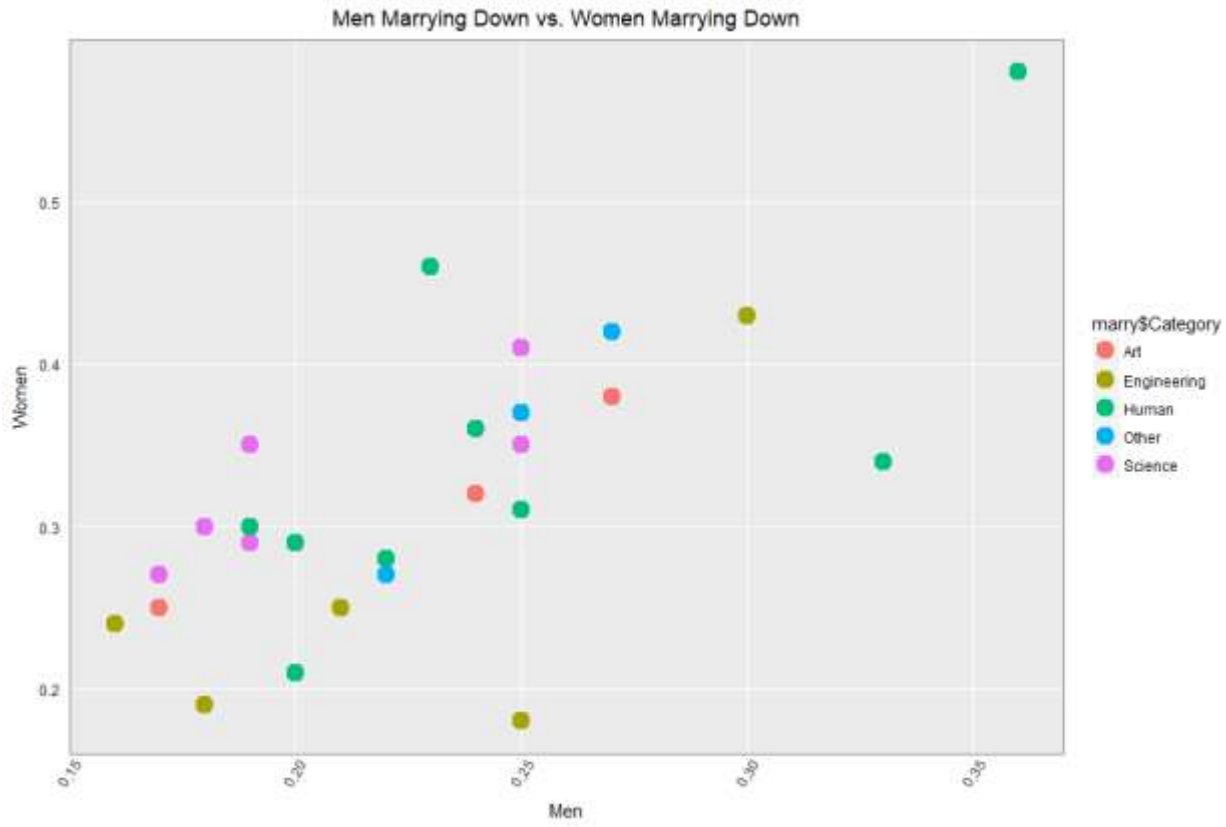


Figure 2: Scatter Plot of Gender Married-Down rate by Major Category

axes.

With the information from **figure 2** we now have two hypothesis for later analysis. The first is that women have a large mean marry-down rate than men. The other is that majors within the same category behave the same. With this in mind we can investigate the first hypothesis.

Below in **figure 3** are several box plots representing the means of both genders in marrying-down, created with ggplot. I overlaid the points corresponding to all of the majors for each gender to give the viewer a good idea of what the distribution of points really look like. I also split both genders by their degree types in order to exam the

distribution of these as well. We see that there is an obvious difference between the means of women and men, furthermore, the variation of points for women is much greater than that of men. For both degree types under the gender of men the boxes are shorter than that of the women. We also see that for both genders the BA's mean is slightly greater than that of its counterpart. However, the distribution of the Bachelor of Science box plots have a greater distribution.

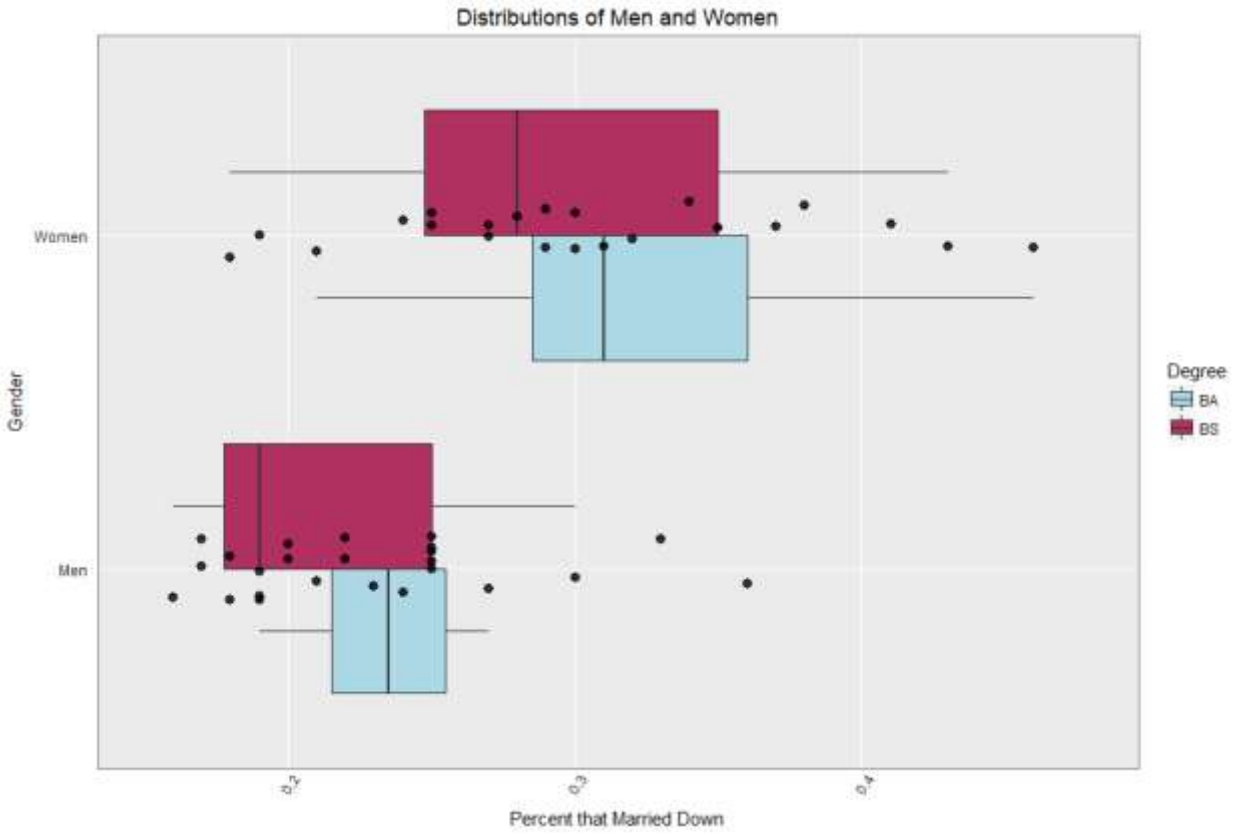


Figure 3: Boxplots of Gender Married-Down Rates by Degree Type

From this plot we can conclude that men have a noticeably smaller distribution and mean rate marrying down than women. We can also conclude that BA's have a larger mean but smaller distribution than BS's.

Now that a lot of the categorical variables and the marry-down rates have been examined I wanted to create a visualization to exam the matrix of majors and their marriage scores. With 17 variables on both axes any graph would be in 17 dimensions. A singular set of data of this magnitude required a special sort of analysis. Keeping the shape of the table I took the marriage scores and applied it to a heat map (**Figure 4)**.

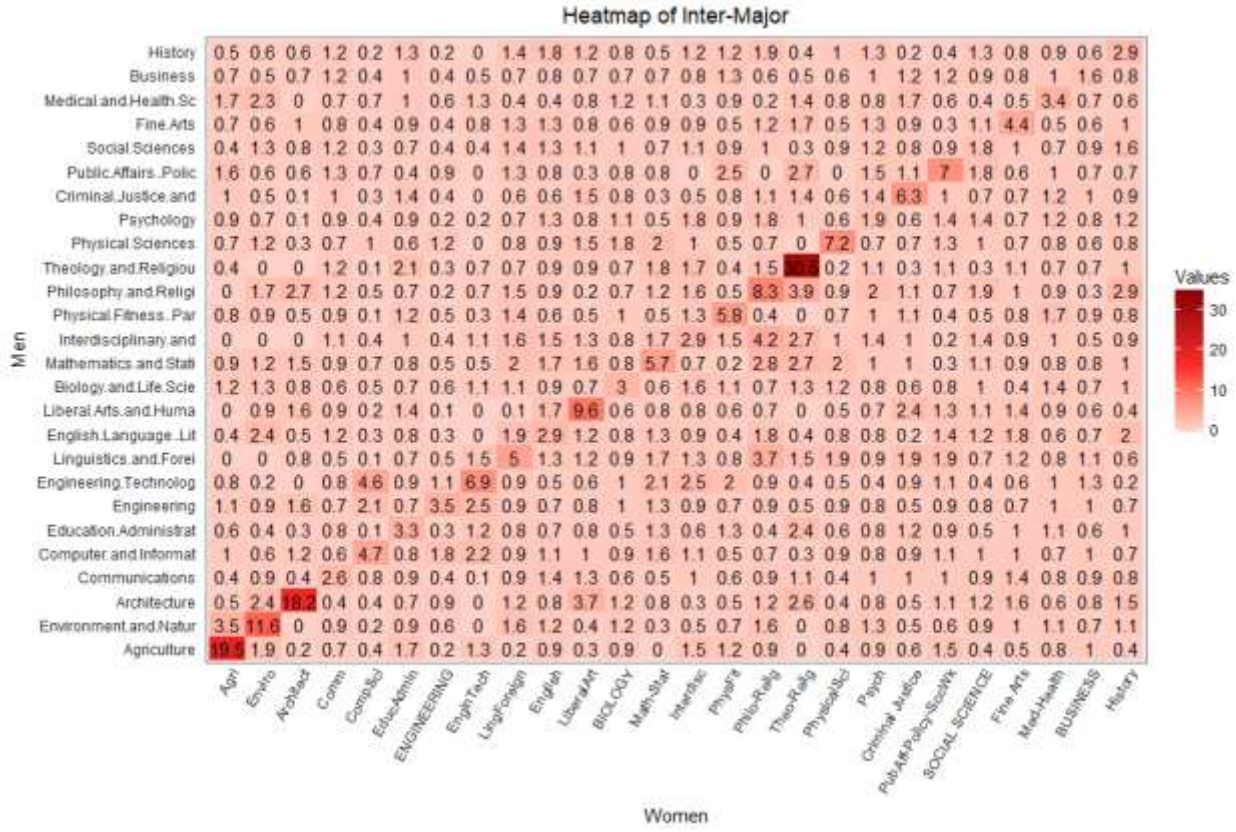


Figure 4: Filled Heat Map of Marriage Scores

The heat map is aesthetically pleasing and is an interesting way to view individual points of data. Overlaying the marriage scores of each space also makes it interesting for random observor. This gave the data set the discovery value that would be lacked in other analysis.

In **figure 4** we see the scale ranges from 0 to 30 with the largest being the darkest red. There was also the challenge of choosing the correct midpont for the

transition of colors. Surprisingly enough the best midpoint was in the middle at 15, even though a majority of the values are below 15. What this allowed though was to focus in on the overbearing pattern. Right through the diagonal of the heatmap there are continuously dark values. This coincides with matching majors having the highest marriage score for their individual majors in every single major. While the results aren't necesarrily surprising it is interesting to see how blantant it is for some majors. For majors like theology/religion, achitecture, agriculture, and liberal art their intra-major marriage rates are much higher than the values within their own majors and within the the other intra-majors marriage rates. Intuition can be applied in these circumstances to hypothesize why these majors have such high rate. For the majors of theology/religion it could be assumed that those with a regligious affinity of that level seek like minded people. The dominance of rural land for agriculture could also explain why those within this major marry. However, majors like architecture and liberal art lack the cultural bias that the others have. This relationship shows room for further analysis of these behaviors.

With this information, that majors have a great intra-major marriage rate, it would make sense that the categories in which majors are in would also share this behavior. To exam this I used ggplot to create a bar graph and 5 facets, each showing the total marriage scores for each category. The bar graphs were faceted by the categories themselves ultimately showing how each category married the other categories. The faceted bar graph in **figure 5** shows us what we were expecting on the scale of categories, categories prefer to marry each other. While the humanities overbear every category due to their population the intra-category levels for every category are either

the highest or the second highest right under humanities. What this means is that the trend found in our last graph is continued into the larger categories. Also, for both engineering and science the intra-category marriage rates are the highest. This behavior is reminicent of our first graph where these two majors had lower marry-down rates. Could that be due to the fact that they strongly intra-major marry?

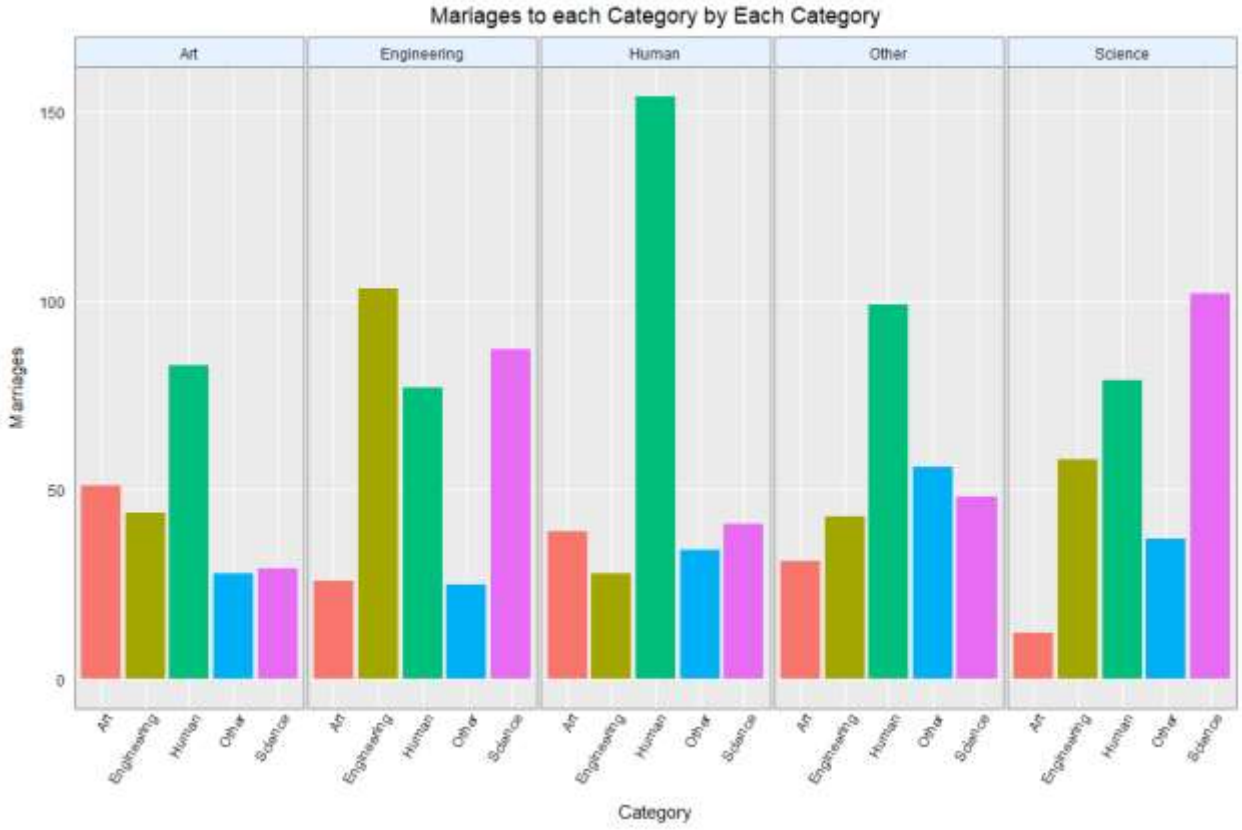


Figure 5: Marriage Rates of Categories by Categories

As mention before, I selected this data set in order to try and use clustering to extract meaningful conclusions form the data. In order to do this I used a package called “facetoextra” and “cluster”, the two packages combined allowed me to easily create a cluster graph. The first thing I did before creating the graph was create a dendrogram to try and investigate the behavior of the clusters. I assigned the results of the hcut function on the data for the married down variables for both genders, to a variable. I

then used that variable in the fviz_dend function in order produce the dendrogram in **figure 6**. In the dendrogram we see 4 distinct groups, all of which do not show conclusive results. Especially at the lower heights majors from the same category often do not group together. It isnt until after a hieght of 1 that intra-category groupings start

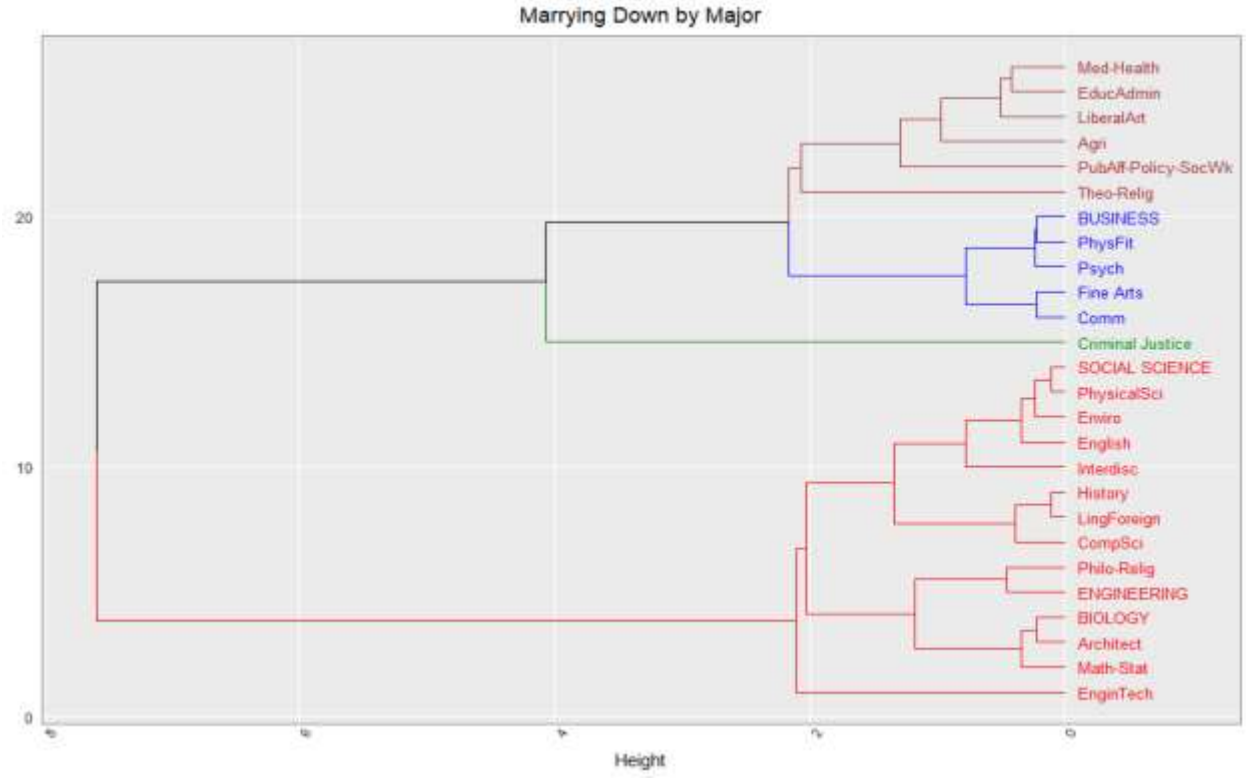


Figure 6: Dendrogram of Marrying Down Rate

to happen. But this could prove to be an even more important finding. Maybe if we were to exam these relationships over time we could see if the same trend still emerged, then we could work toward a conclusion that majors such as social science and physical science often show the same behaviors because their connecting height is so low. Also a key point to notice in this dendrogram is that Criminal Justice is without a group until a height of 4.

Before I began graphing the clustered data I started with a method used to examine how many centers I should specify. This method is called the silhouette

method and it used the data I was going to use for clustering. It then outputs a level of average silhouette widths for each number of cluster centers (**figure 7**). Reading this graph we see that the optimal number of centers is two.

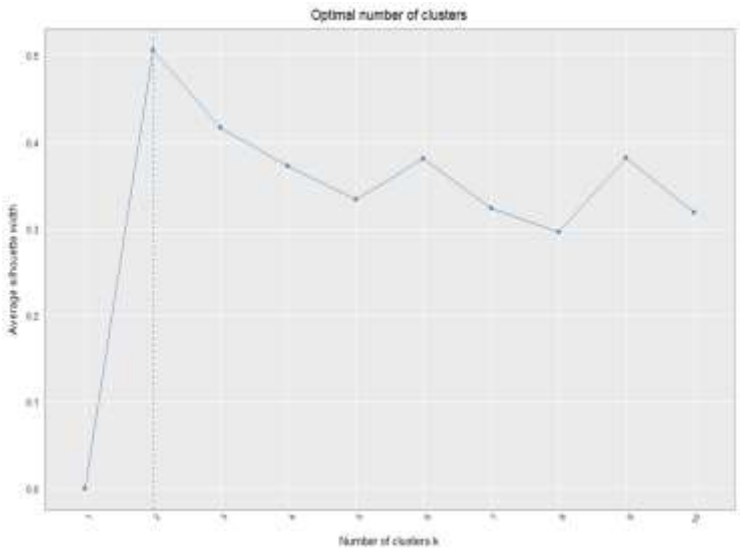


Figure 7: Silhouette graph

The cluster graph is displayed in **figure 8,** where the two clusters populate most of the graph area, the two centers are signified by the red circle and blue triangle. Within each cluster again it is hard to see a trend that corresponds with previous findings. Clusters aren't limited to categorical variable but instead hold a large amount of all categories. As stated before

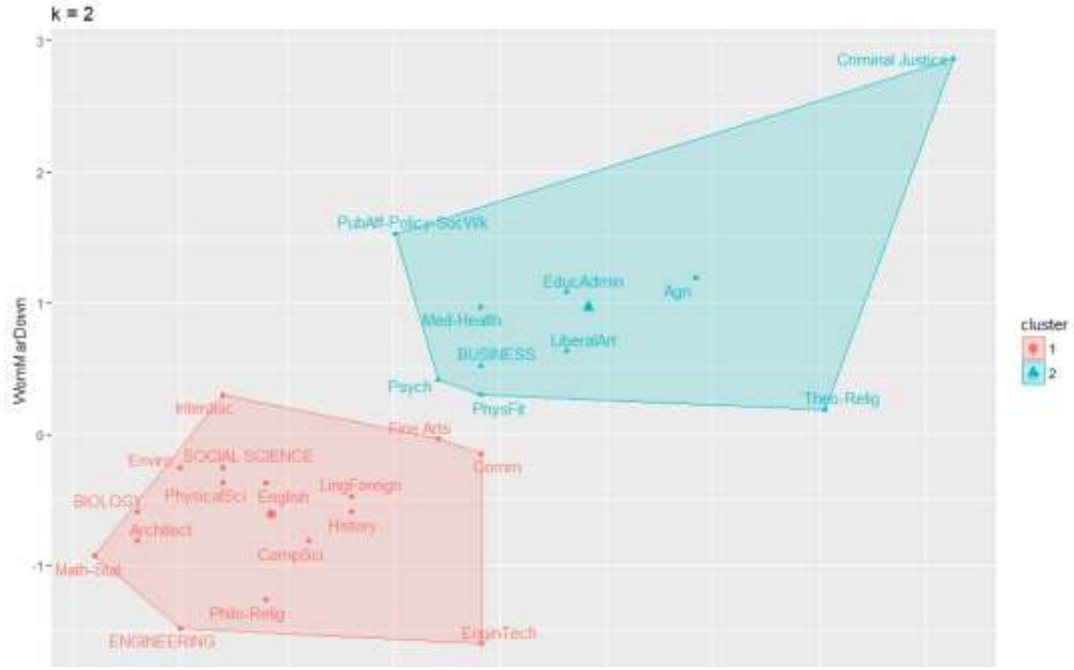


Figure 8: Cluster Graph of Marrying Down Rates

these clusters could prove useful in a time series if these trends continue. They could be used to narrow down what type of major a certain marry-down rate corresponds to. Other than the hypotheticals it is tough to take useful conclusions from this graph. For demonstration purposes I added a graph with three clusters above (**figure 9**), this shows that the silhouette diagram proved useful and, like the dendrogram, criminal justice majors are an outlier.

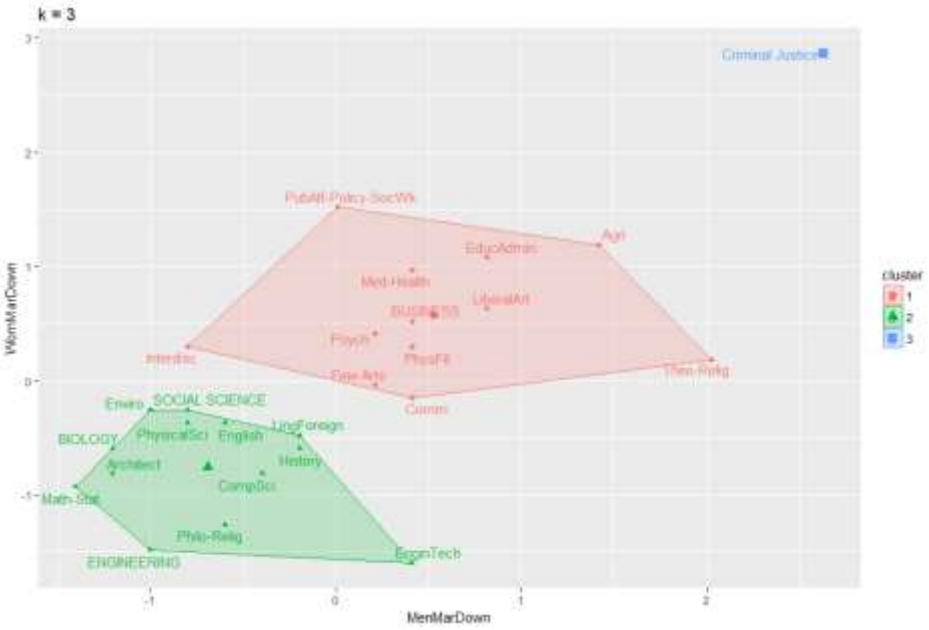
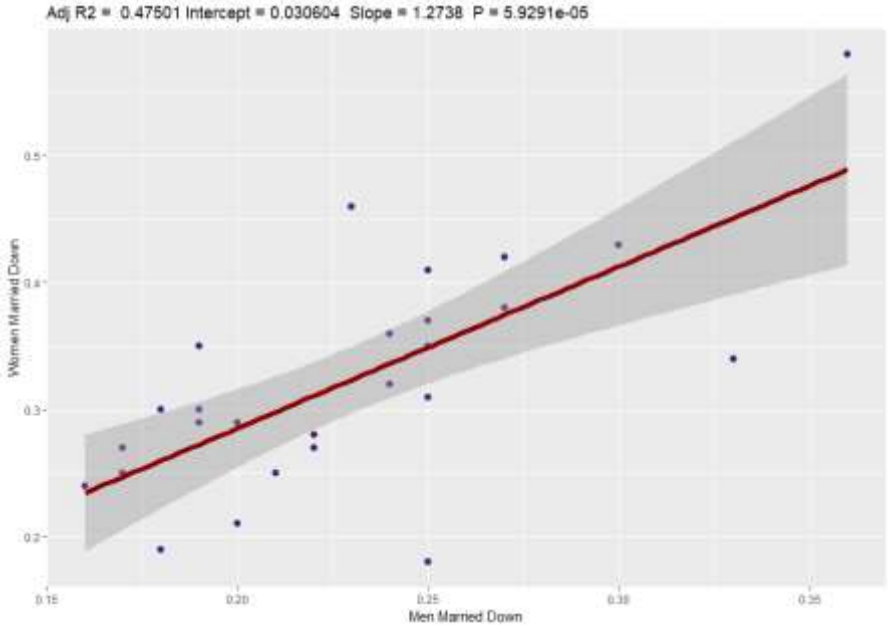


Figure 9: A Bad Clustering Graph

My final graph uses this same plot of data as above (**figure 8**) in an attempt to create a linear regression model to predict a gender's marry down rate from the opposite gender's rate. Using the linear model function I modeled women by men. I then used a function to display summary statistics with the linear model. Below we see the regression



modeled with a p value of .00005, this value is lower than a p value of .01 thus we can accept this graph. The R-squared adjusted value is not in our favor at .48 but is still substantial for analysis. The slope and intercept together give us a linear equation of y = 1.27x +0.03 where the y value is the rate of women who married down and the x value is the rate of men who married down.

Appendix: Full R Code

```
library('ggplot2')
library('gridExtra')
library('ggrepel')
library('reshape2')
library('tidyverse')  # data manipulation
library('cluster')    # clustering algorithms
library('factoextra')
library('ggrepel')


source("C:\\Users\\David\\Downloads\\STAT 463\\Final Project\\hw2.R")

marry = read.csv("C:\\Users\\David\\Downloads\\STAT 463\\Final
Project\\Finaldata.csv", header = T, sep =",", stringsAsFactors = F, dec=".")

#########################################################
#lets see if there is obvious trend between marrying down and major
ggplot(marry, aes(x = MenMarDown, y = WomMarDown)) +
  geom_point(aes(MenMarDown, WomMarDown, fill = marry$Category, color =
marry$Category, shape = marry$Degree), size = 4, stroke = 1.2) +
  labs(title = "Men Marrying Down vs. Women Marrying Down", x = "Men", y =
"Women") +
  hw2
#No but we notice that women may more than men

#########################################################

forbox2 = melt(marry, id.vars = "Degree", measure.vars = c("MenMarDown",
"WomMarDown") )
forbox2 = subset(forbox2, Degree != "Either")
#Lets make a boxplot to see this


ggplot(forbox2, aes(x = variable, y = value, fill = Degree)) +
  geom_boxplot( outlier.shape = NA) +
  geom_point(aes(fill = variable), size = 2.5, shape = 19, alpha = .8, position =
position_jitterdodge(), show.legend = F) +
  labs(title = "Distributions of Men and Women", x = "Gender", y = "Percent that
Married Down") +
  hw2 +
  coord_flip() +
  scale_fill_manual(name = "Degree", values = c("lightblue","maroon",NA,NA), limits =
c("BA","BS")) +
  scale_x_discrete(labels=c("Men", "Women")) +
```

```
  scale_y_continuous(limits = c(0.15,0.48))

#############################################################################
###########

g = melt(table)
ggplot(g, aes(Var1, Var2)) +
  geom_tile(aes( fill = value )) +
  geom_text(aes(label = round(value, 1))) +
  labs(title = "Heatmap of Inter-Major", x = "Women", y = "Men") +
  scale_fill_gradient2(low = "white", high = "red4", mid = "firebrick1",
                       midpoint = 15, limit = c(0,35),
                       name="Values") + hw2

#############################################################################
###########

table2 = cbind(marry$Category, table)
colnames(table2)[1] = "Category"
table2 = data.frame(table2)
bycat = aggregate(. ~ Category, data  = table2, sum)
colnames(bycat)[2:27] = marry$Category
try = t(bycat)
try = try[-1,]
try = cbind(marry$Category, try)
colnames(try) = c("Category","Art","Engineering","Human","Other","Science")
try = try[,-7]
#Because there were so many Humanities we arrange the data
try = data.frame(try)
try2 = aggregate( . ~ Category, data = try, sum)
try2 = try2[,-1]
try2 = cbind(c("Art","Engineering","Human","Other","Science"), try2)
colnames(try2)[1] = "Category"
try3 = melt(try2, id.vars = "Category")


ggplot(try3, aes(x = Category, y = value)) +
  geom_bar(aes(fill = try4$Category), stat = "identity") +
  labs(title = "Mariages to each Category by Each Category", x = "Category", y =
"Marriages") +
  hw2 + facet_grid(~ variable) +   guides(fill = F, colour = "none")


#############################################################################
3
```

```
marrydowns = marry[-6]
marrydowns = marrydowns[-4]
marrydowns = marrydowns[-1]
marrydowns = marrydowns[-5:-32]
marrydowns = marrydowns[-1:-2]
row.names(marrydowns) = marry$Major

res = hcut(marrydowns[1:2], k = 4, stand = T)
fviz_dend(res, rect = F, cex = 0.7, k_colors = c("red","green4", "blue", "brown"), horiz =
T) + hw2 + ggtitle("Marrying Down by Major")

########
#https://uc-r.github.io/kmeans_clustering#elbow


fviz_nbclust(marrydowns[1:2], kmeans, method = "silhouette") + ggtitle("Marrying
Down by Major") + hw2

big2 = kmeans(marrydowns[1:2], centers = 2, nstart = 25)
fviz_cluster(big2, data = marrydowns, geom = c("point", "text"), repel = T, ellipse.type =
"convex", ellipse.alpha = 0.2) +  ggtitle("k = 2") +
  hw2

big3 = kmeans(marrydowns[1:2], centers = 3, nstart = 25)
fviz_cluster(big3, data = marrydowns, geom = c("point", "text"), repel = T, ellipse.type =
"convex", ellipse.alpha = 0.2) +  ggtitle("k = 3") +
  hw2


#####

fviz_nbclust(table, kmeans, method = "silhouette")

matclus = kmeans(table, centers = 2, nstart = 40)
fviz_cluster(matclus, data = table,stand = T, geom = c("point", "text"), repel = T,
ellipse.type = "convex", ellipse.alpha = 0.1) +  ggtitle("k = 2")


##############################################33


ggplotRegression <- function (fit) {

  require(ggplot2)

  ggplot(fit$model, aes_string(x = names(fit$model)[2], y = names(fit$model)[1])) +
    geom_point(col = "purple4", size = 2.5) +
```

```
   stat_smooth(method = "lm", col = "red4", size = 2) +
   labs(title = paste("Adj R2 = ",signif(summary(fit)$adj.r.squared, 5),
                "Intercept =",signif(fit$coef[[1]],5 ),
                " Slope =",signif(fit$coef[[2]], 5),
                " P =",signif(summary(fit)$coef[2,4], 5)), x = "Men Married Down", y =
"Women Married Down")
}
linear = lm(marry$WomMarDown~marry$MenMarDown)
summary(linear)
ggplotRegression(linear)
```

Data Set Extracted From: https://osf.io/h2bny