Multivariate Analysis Assignment 3

Zane Hassoun - ID: 210031934

Group 5

April 4th, 2022

MT 5758

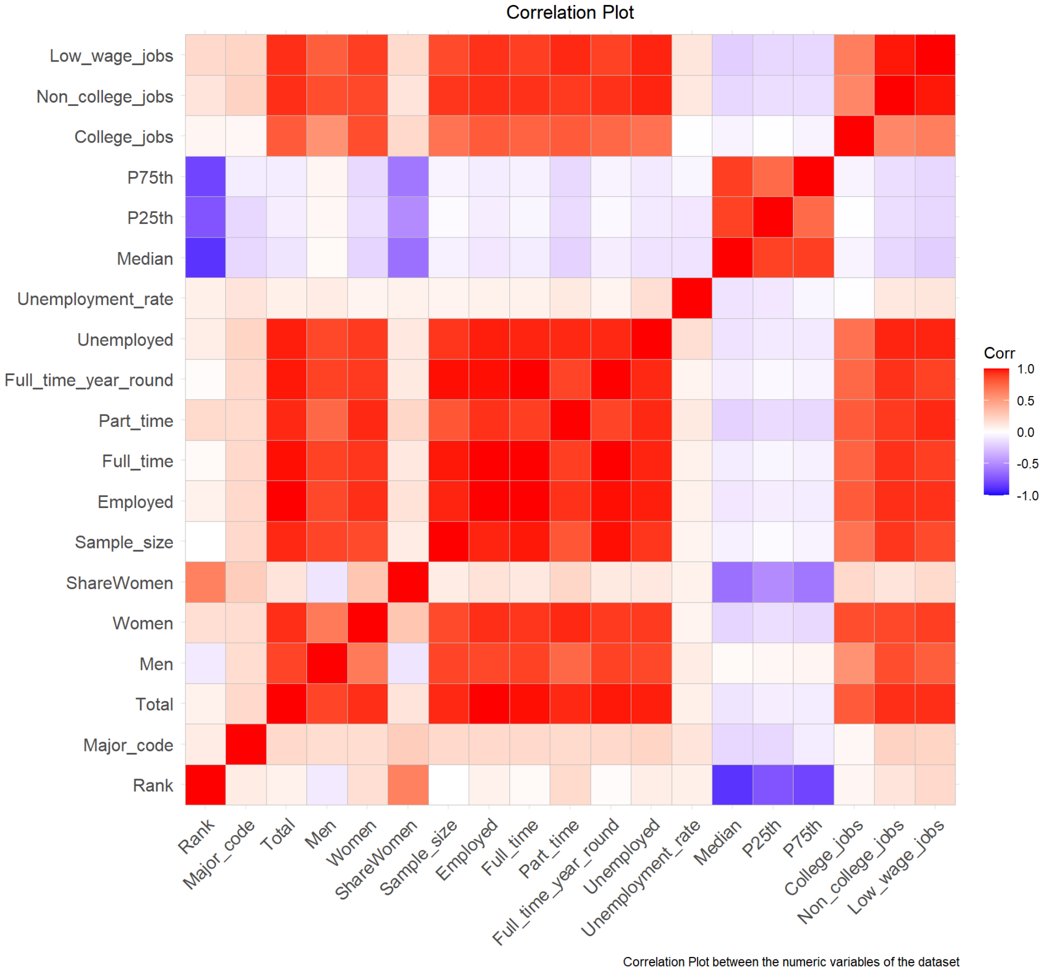
Introduction

For 17–19-year-old students, a common problem arises when they begin to apply to university: in what course are they to undertake further study? Generally, there is some guidance through university counsellors, parents, peers, or mentors, but there is a swath of information available and it is difficult to identify pertinent information and thus determine the most optimal route to university entry . Beginning to define ”optimal” opens the door to significant nuance: Is the student independently wealthy? Do career outcomes matter? Do they value money, over what they enjoy? Do they have moral underpinnings that direct them to take on work that is meaningful? The list is ad-infinitum. In an attempt to assist in determining a solution, this analysis seeks to propose a guide for students that replaces the typical ”Major Category” utilized to begin one’s search (i.e. Engineering, Physical Science, Arts, Humanities etc) and utilizes instead, clusters of degrees based on different post-graduation employment statistics. For each specific university major, a cluster of similar majors will be generated based upon the statistics. The compelling aspect of this analysis is the ability to partition the data based upon a larger number of variables than a comparison of median salaries or unemployment statistics. Once completed the student will be able to make informed decisions with numerous options within and across clusters. Given that the perception of data above three dimensions all but impossible, the challenge of this analysis is to reduce the data dimensions while trying to retain as much of the variation within the data as possible. Fundamentally, I aim to derive an algorithm that can partition the data to group university degrees based upon the various available employment statistics. The motivating example is as follows: a young woman enters her college counselling office and states she is somewhat interested in Math and Physics, but it is integral to her life that she finds a degree that offers her excellent employment opportunities post university, with the caveat that she would prefer not to be in a completely male dominated profession. What avenue should she take? The clusters would give guidance on that answer. The first step would be to identify the cluster that fulfils the employment and percentage of women and to determine if any viable majors exist; If not, the procedure would involve moving to the next cluster that has the best compromise until the student reaches the most favourable opportunity is determined. Ideally there are enough options in the first cluster of degrees that enables the students to achieve their goal.

Methods

The data set the work will be carried out upon contains 173 different university majors, that are grouped into 16 different major categories (Engineering, Education, Law, etc.). This data set provides wage data, share of women, employment status in addition to individual major sample sizes. These data are quite clean, and to prepare for analysis, the column ”Major Category” was mutated into a factor variable rather than a string (text).For exploratory data analysis, I created three different groups based on median income: Low, Medium and High, partitioning the degrees fairly uniformly to get an initial visualization of the groupings based on median income. Exploratory data analysis was then conducted to determine the datatype and distribution of each feature. To effectively reduce the dimensions of the data, many of the variables need to be correlated. These data were, in fact, highly correlated, and therefore it appeared advantageous to apply Principal Component Analysis (”PCA”) to the data. PCA looks to reduce the dimensions of the data by orthogonally projecting the data points onto lower dimensional axes with hopes of retaining as much variability in the data as possible. Since humans can only visualize data at best in three dimensions, but more usefully in two dimensions, I needed to determine how much of the total variance was explained in the first two (think of and x and y scatter plot) or three (an x,y,z plane) principal component axes. The high levels of variance within the first three components make plotting the data in low dimensions attainable, and therefore, using those axes to obtain and visualize clusters is deemed appropriate. To obtain those clusters, I used the derived PCA axes in place of the original data and applied two different types of clustering: K-Means, and Hierarchical clustering. K-Means places K (specified) different centroids within the data set and iteratively searches for an optimal solution that reduces the Within Cluster Sum of Squares i.e., the distance between each data point and its respective cluster centre. Hierarchical clustering can take different algorithms: Ward, Single, Average and Complete linkage, which are agglomeration methods with various equations (see appendix). Each method was reviewed to evaluate the best option. With this new clustered data, the goal was to replace the Major Category column of the data set with the new employment statistics based clusters. Since there were 16 major categories, I decided to choose 15 total clusters, with which to implement a clustering algorithm. I chose 15 rather than 16 because the category called ”interdisciplinary”, would likely have yielded many clusters across different majors when compared to employment data. Therefore I made the decision to omit this ”catch-all” category.

Figure 1

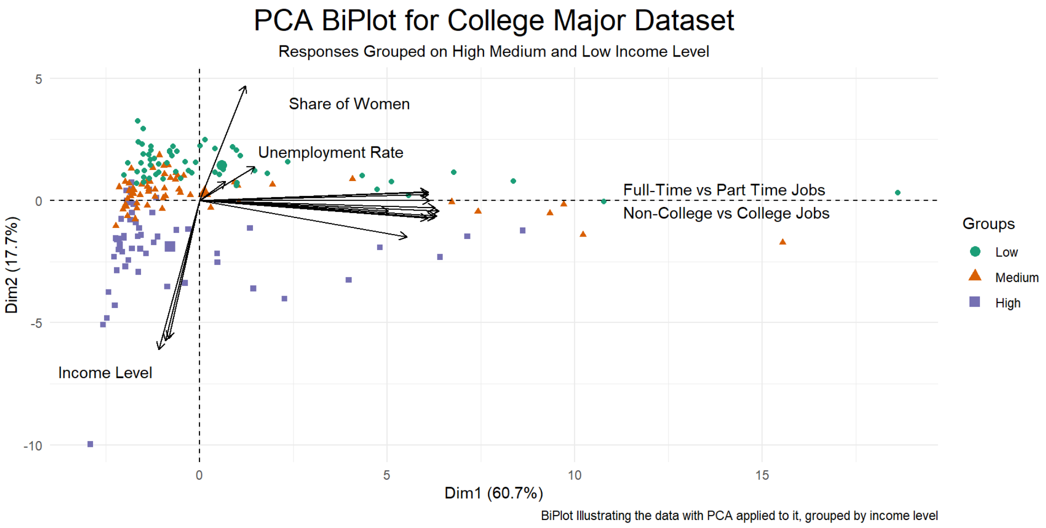


Results

Seen in figure 1, the data set contains features that are highly correlated. Specifically, more than half of the feature pairs have greater than 90% correlation. There were expected correlations: median 25th and 75th income quantiles, but also less than expected ones: large inverse correlation between income levels and share of women. While great levels of correlation existed, it was encouraging to see there were groups of features that are uncorrelated, which proved to be advantageous for this project as I sought to reduce the not only the number of dimensions, (which correlation helps with) but also to cluster different variables, which is also necessary for variability between the groups.

Once Principal Component Analysis had been applied to the data, the evaluation of variation explained by each principal component needed to be evaluated. From the component projections, it was shown that almost 80% of the variance was explained in the first three components and therefore encouraged the research to continue.

Figure 2



With the new component axes figure 2 was created to try to determine if the associations between features and data points was well defined. The biplot showed there were 3 major directions: share of women and unemployment rate; income level, and types of jobs. This plot segmented the data by an initial grouping of Low, Medium, and High income. We will return to this plot when the new clusters are defined. From the comparisons of different clustering methods, figure 3 and 4, it was apparent that K-Means clustering created the most well-defined groups that did not leave individuals like Single and Average linkage did. This is also well illustrated in figure 4, a silhouette plot that shows the width of the clusters and illustrates any deviation from within their cluster into another. From examining both plots it became evident that K-means was the best available method for the clustering. Aside from Ward linkage, the hierarchical methods did not yield very discernible results, and culminated in very discontinuous groups.

Figure 3

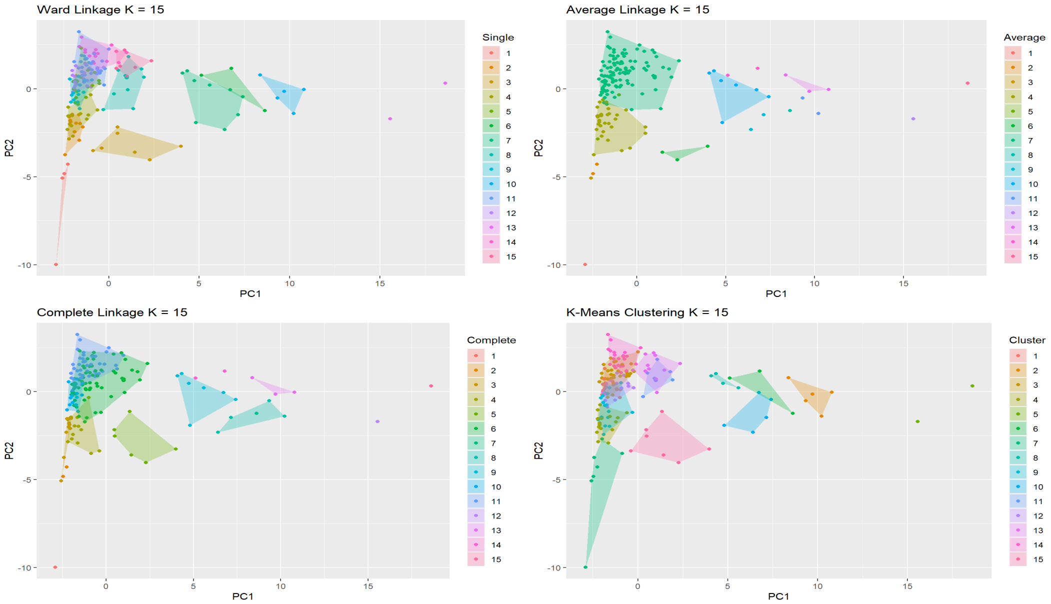
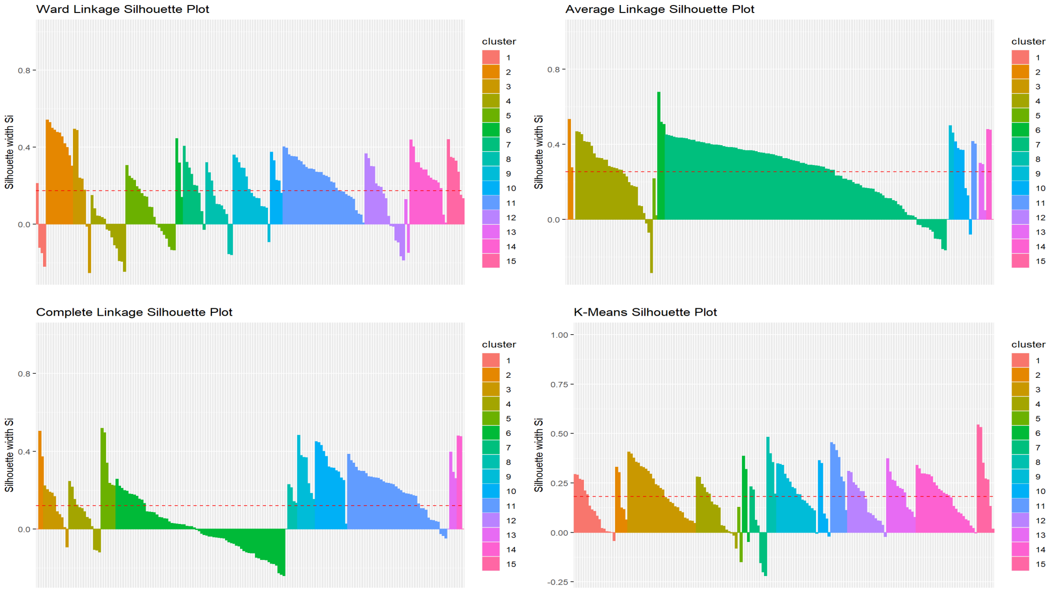
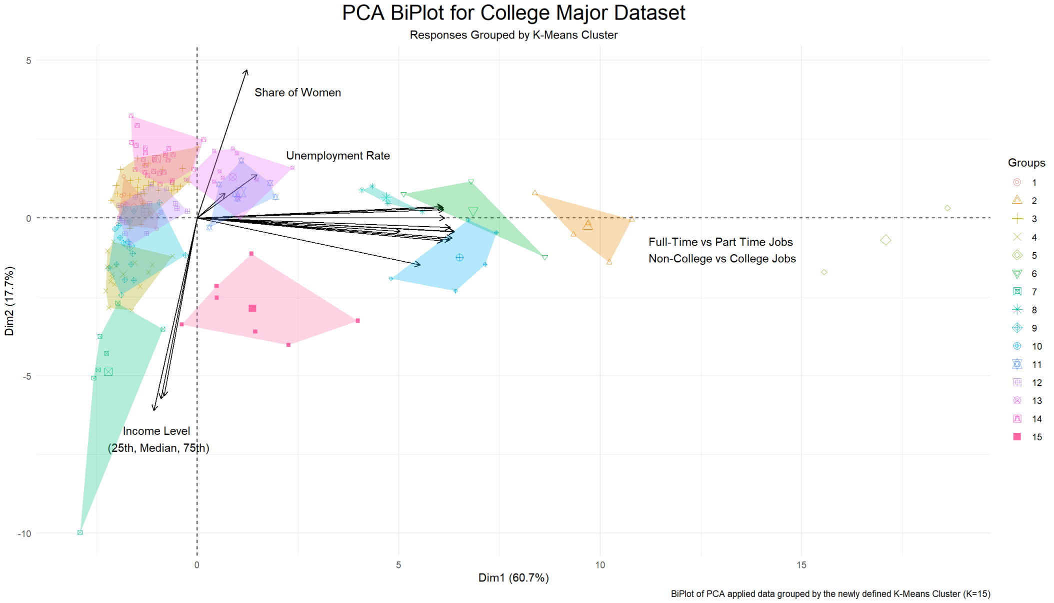


Figure 4



I proceeded with K-Means which yielded these different groups. The figure details the new coherence across the different groups that conforms better with generally accepted American ideals of college majors. I imposed these new well-defined clusters, onto the biplot in figure 5 to obtain a visual aid to help determine the features of each cluster. The plot shows clusters with a larger value on PC1 (x-axis) are positively associated with higher numbers of jobs, where clusters with larger values on PC2 (y-axis) are positively associated with higher percentage of women, as well as lower wages. Since the new clusters have variation across both axes, the algorithm can successfully recommend 15 clusters of majors that vary across the available employment statistics.

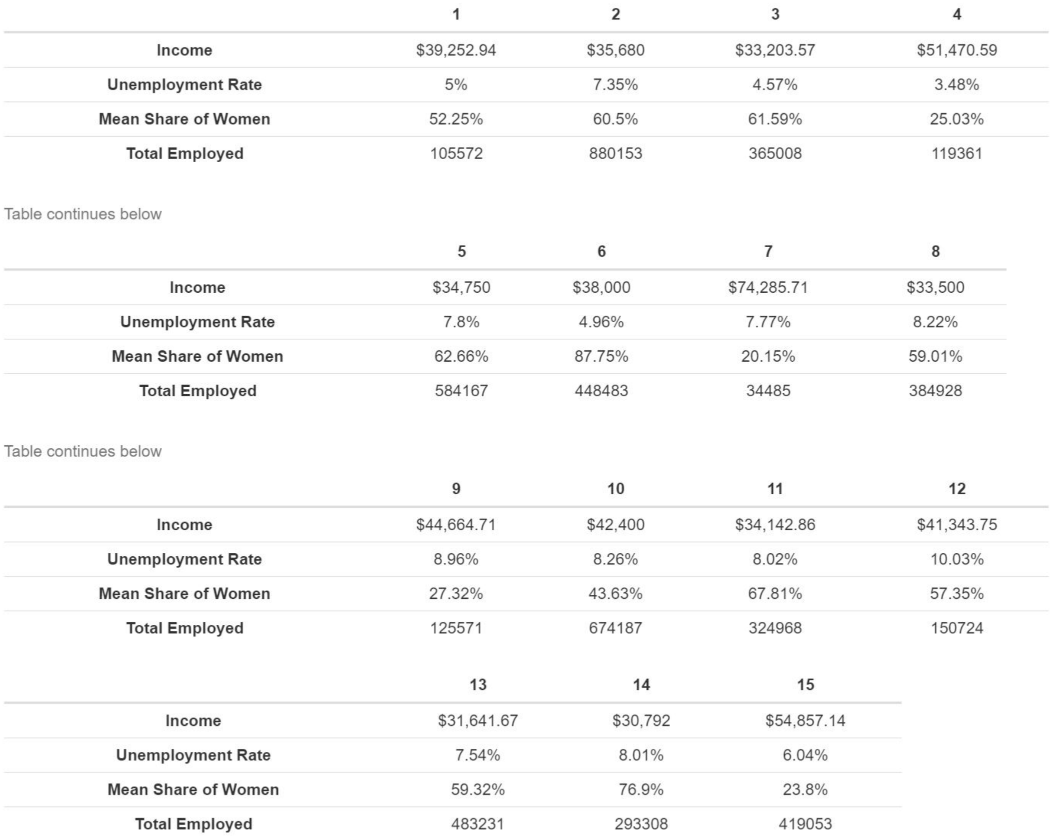
Figure 5



Discussion & Conclusion

This analysis proved to be fruitful and successful. The methods were aptly applied to the data set and yielded results that solved the proposed problem of replacing the traditional "major" category utilized by college counsellors. Replacing this category with an employment statistics based clustering algorithm. The findings in figure 5 effectively detail the groupings that contain variability between and across features, and when comparing the means in the table in figure 6 distinct differences are seen across the feature space. The decision to combine Principal Component Analysis in addition to K-Means Clustering proved to be very effective, addressing the initial constraint of visualizing the clusters across 17 dimensions, while still retaining a high percentage of variability within the data. In fact, many of the clusters were non-overlapping and explored almost the entirety of each plotted principal component dimension. Additionally, I was pleased to see there were movements across the typical question: what has the highest salary? I believe this analysis has addressed the motivating example discussed in the introduction and gives plenty of options for a student to contemplate, allowing them to reach an informed decision before embarking upon their university/academic pathway. If building upon these findings in future projects I would have moved forward with these data to and included more information around the domains of the topics. I would also have utilized Natural Language Processing to synthesize this information. Considering for example: environmental focus or money focus or children or people skills etc I believe it could be a powerful addition to help with clustering for the students. I am certain that this could be a very useful tool not only for high school college guidance counsellors, but also for universities to group their degrees differently. Perhaps they would find utility in creating clusters of groups to market the degrees.

Figure 6



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Income	$39,252.94	$35,680	$33,203.57	$51,470.59	$34,750	$38,000	$74,285.71	$33,500	$44,664.71	$42,400	$34,142.86	$41,343.75	$31,641.67	$30,792	$54,857.14
Unemployment Rate	5%	7.35%	4.57%	3.48%	7.8%	4.96%	7.77%	8.22%	8.96%	8.26%	8.02%	10.03%	7.54%	8.01%	6.04%
Mean Share of Women	52.25%	60.5%	61.59%	25.03%	62.66%	87.75%	20.15%	59.01%	27.32%	43.63%	67.81%	57.35%	59.32%	76.9%	23.8%
Total Employed	105572	880153	365008	119361	584167	448483	34485	384928	125571	674187	324968	150724	483231	293308	419053

References

[1] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2021

[2] James et. al. An Introduction to Statistical Learning with Applications in R, 8th Edition, 2017

MVA_Project_ASN3

Zane Hassoun

3/27/2022

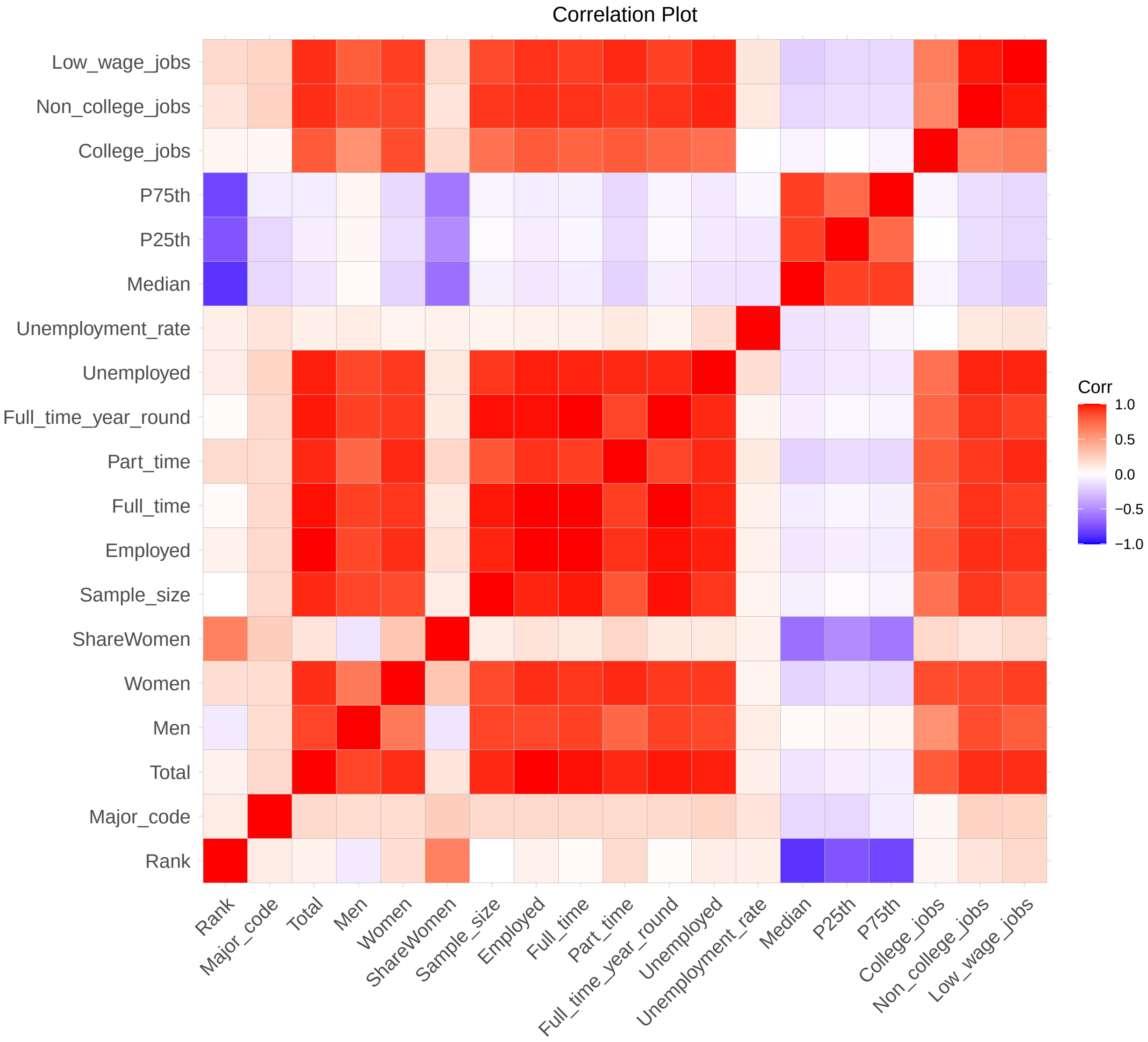
Some Exploratory Data Analysis The deviation of Dollars and Cents is not particularly of material interest so I will try and break up some groups of data into high medium and low income levels. The hope is that k-means or another clustering algorithm can cluster across these not the same ones

```
dta_factored = raw_dta %>%
  mutate_if(is.integer, as.numeric) %>%
  mutate(Income = ifelse(Median <35000, "Low", ifelse(Median < 41000, "Medium",
                                                      "High")))%>%
  na.omit()
dta_factored$Income = factor(dta_factored$Income,
                             levels = c("Low", "Medium", "High"))
pander::pander(table(dta_factored$Income))
```

Low	Medium	High
60	57	55

The intuition is that there is a lot of correlated data at hand here - so we will get a correlation plot

```
## ggplots version of a correlation plot - removed non numeric variables.
ggcorrplot(cor(dta_factored %>%
      select(-Major, -Major_category, -Income) %>%
      na.omit())) + labs(title = "Correlation Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(caption = "Correlation Plot between the numeric variables of the dataset")
```



Correlation Plot between the numeric variables of the dataset

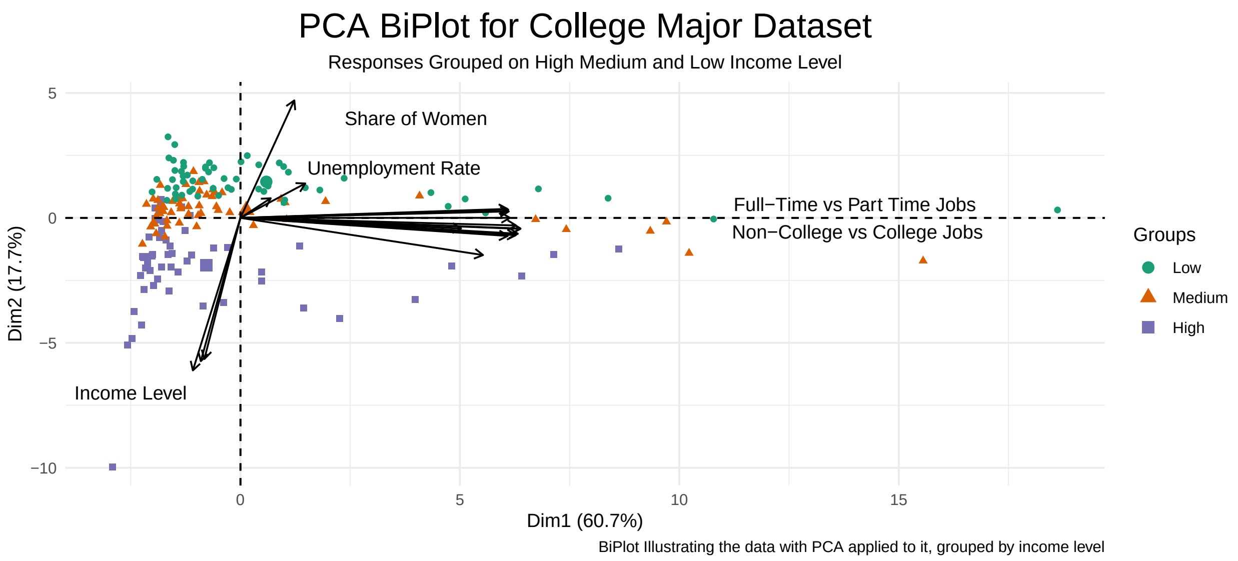
Clearly a lot of correlation between a large chunk of the variables so PCA is a viable candidate

```
#Since the values here are not all of the same quantity i am going to scale
#and center the varables so they are of equal weight in the PCA decomposition

#This is then using the r-base pca algorithm and obtaining
#the relevant information

X = scale(dta_factored %>%
      select(-Major, -Major_category, -Income, -Rank) %>%
      na.omit(),
      center = TRUE,
      scale = TRUE)
```

```
PCA_Model = prcomp(X)
scores = as.data.frame(as.matrix(X)%*%PCA_Model$rotation)
eigenvalues = PCA_Model$sdev**2
```

```

#Using that information - creating some cumulative
#variance plots and an itial biplot
vars_exp = round(eigenvalues/eigenvalues,2)
cs = round(cumsum(eigenvalues)/sum(eigenvalues),2)
prop_exp = data.frame("PC" = c(1:18),
                      "Raw" = vars_exp,
                      "Cumulative" = cs )
var_exp = ggplot(data = prop_exp) +
  geom_point(aes(x = PC,
                 y = Cumulative),
             size = 4)+
  labs(title = "Cumulative Variance Explained",
       x = "Principal Component Index",
       y = "Proportion of Variance Explained",
       caption = "Plot illustrating the proportion of variance explained by the Principal Componenet Ana
  theme(plot.title = element_text(hjust = 0.5))


#This is the biplot with various labels
biplot_pca = fviz_pca_biplot(prcomp(X),
                devcol.var = 'black',
                # var.scale = 0,
                # alpha = 0.5,
                label = TRUE,
                habillage=(dta_factored$Income))+
                # addEllipses = TRUE)+
  scale_color_brewer(palette="Dark2")+
  theme_minimal()+
  annotate('text', label = "Share of Women", x = 4, y = 4)+
  annotate('text', label = "Income Level", x = -2.5, y = -7)+
  annotate('text',
           label = "Full-Time vs Part Time Jobs\n Non-College vs College Jobs",
           x = 14, y = 0)+
  annotate('text', label = "Unemployment Rate", x = 3.5, y = 2)+
  labs(title = "PCA BiPlot for College Major Dataset",
       subtitle = 'Responses Grouped on High Medium and Low Income Level',
       caption = "BiPlot Illustrating the data with PCA applied to it, grouped by income level") +
  theme(plot.title = element_text(hjust = 0.5, size = 20),
        plot.subtitle = element_text(hjust = 0.5))
biplot_pca
```

Principal Component Analysis



We capture a lot of the variance almost 80% by two components and 90% by four so it looks good to continue

Clustering

K Means Clustering The first task at hand is to evaluate the number of clusters we want… hypothetically i would like a sizeable number of clusters to use but for interests sake lets see if it partitions the pca data in a similar way to the way the Low Medium and High Groups do

Justification of Clustering The idea is we have a thing called a major category… I instead would like to give you a new category called "Employment Rank" the most employable and worthwhile degress available

This is going to be about the ways that we can deviate from the categories

```
set.seed(1234)
#Viewing unique major categories
unique(dta_factored$Major_category)
```

```
##  [1] Engineering                         Business
##  [3] Physical Sciences                   Law & Public Policy
##  [5] Computers & Mathematics             Industrial Arts & Consumer Services
##  [7] Arts                                Health
##  [9] Social Science                      Biology & Life Science
## [11] Education                           Agriculture & Natural Resources
## [13] Humanities & Liberal Arts           Psychology & Social Work
## [15] Communications & Journalism         Interdisciplinary
## 16 Levels: Agriculture & Natural Resources Arts ... Social Science
```

```
#choosing cluster number
num_clust = 15
#Using r base k-means clustering with 10 different starts

km_clust1 = kmeans(scores,
      center = num_clust,
      nstart = 10)

plot_dta_km1 = data.frame(scores, "Cluster" = as.factor(km_clust1$cluster),
                          "Income" = dta_factored$Income)
```

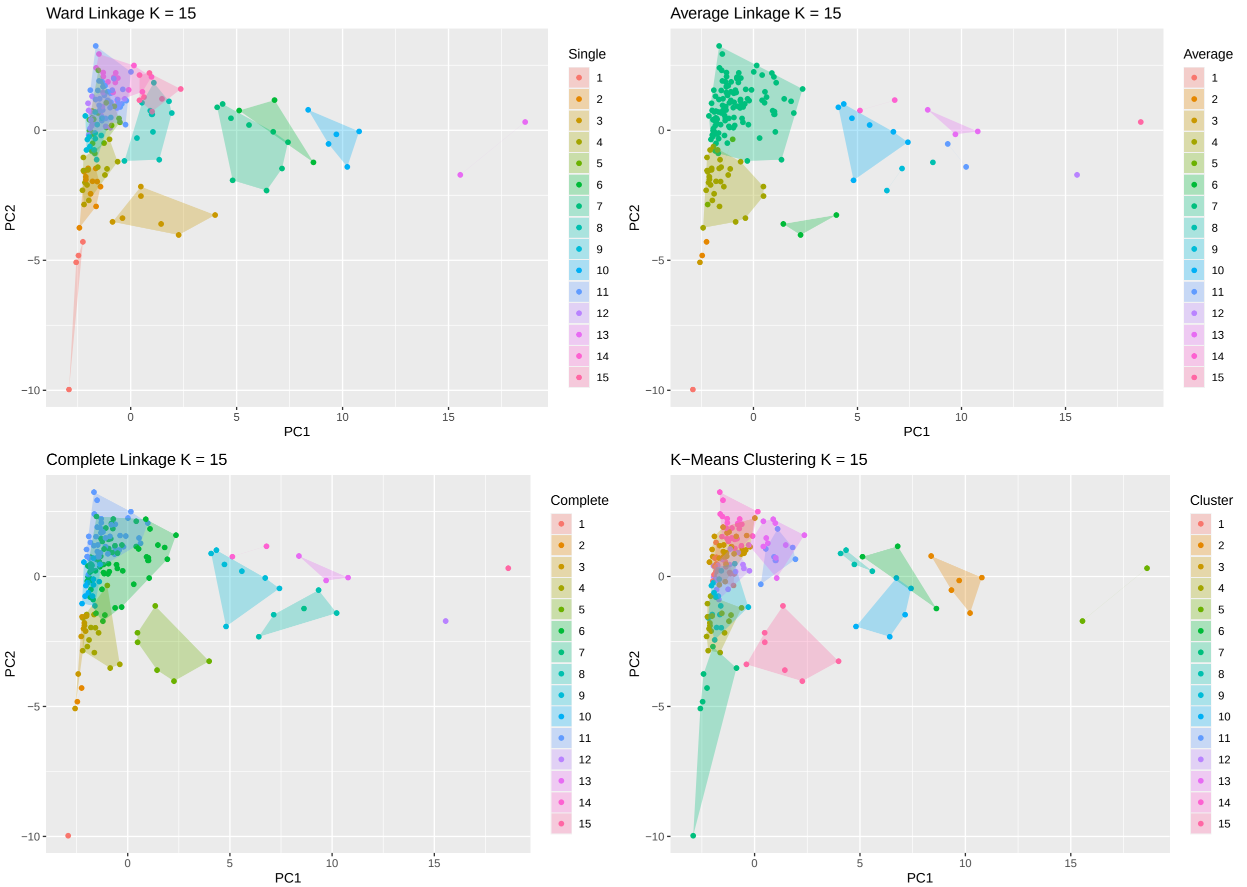
```
kmPlt = ggplot(data = plot_dta_km1) + geom_point(aes(x = PC1, y = PC2, color = Cluster)) +
  geom_convexhull(alpha = 0.3,aes(x = PC1, y = PC2, fill = Cluster)) +
  labs(title = paste("K-Means Clustering K =", num_clust))
```

Heirarchical Clustering For hierarchical clustering I need a distance matrix This chunk of code details the process to obtain the different linkage algorithms available single was omitted - the structure of the data was not worth exploring would create a huge link in top left of biplot

```
pca_dist = dist(scores)

pca_df = data.frame("PC1" = scores$PC1, "PC2" = scores$PC2  )
links = c("average", "ward", "complete")
hClust = c()

for(i in links){
  hClust = c(hClust, cutree(tree = hclust(d = dist(scores), method = i),
                            k = num_clust))
}
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
avg = hClust[1:172]
sing = hClust[173:344]
cmplt = hClust[345:516]

plt_dta_heir = data.frame(scores,
                          "Average" = as.factor(avg),
                          "Single" = as.factor(sing),
                          "Complete" = as.factor(cmplt))

avg_plt = ggplot(data = plt_dta_heir) +
  geom_point(aes(x = PC1, y = PC2, color = Average)) +
  geom_convexhull(alpha = 0.3,aes(x = PC1, y = PC2, fill = Average)) +
  labs(title = paste("Average Linkage K =", num_clust))


sing_plt = ggplot(data = plt_dta_heir) +
  geom_point(aes(x = PC1, y = PC2, color = Single)) +
  geom_convexhull(alpha = 0.3,aes(x = PC1, y = PC2, fill = Single)) +
  labs(title = paste("Ward Linkage K =" , num_clust))


cmplt_plt = ggplot(data = plt_dta_heir) +
  geom_point(aes(x = PC1, y = PC2, color = Complete)) +
  geom_convexhull(alpha = 0.3,aes(x = PC1, y = PC2, fill = Complete)) +
  labs(title = paste("Complete Linkage K =", num_clust))

grid.arrange(sing_plt,avg_plt,cmplt_plt,kmPlt,
             nrow = 2,
             ncol = 2)
```



This code chunk is pulling all the data together to create nice plots with which i compared each clustering output

```
KM_cluster_data = data.frame("Major" = dta_factored$Major ,
                          "Category" = dta_factored$Major_category,
                          "Cluster" = as.factor(km_clust1$cluster),
                          "Income" = dta_factored$Income,
                          "Median_Salary" = dta_factored$Median,
                          "Unemploy_Rate" = round(dta_factored$Unemployment_rate*100,2))

sil_km = silhouette(x = km_clust1$cluster,
                    dist = pca_dist)
sil_sing = silhouette(x = sing,
                    dist = pca_dist)
sil_avg = silhouette(x = avg,
                    dist = pca_dist)
sil_cmplt= silhouette(x = cmplt,
                    dist = pca_dist)

library(factoextra)
km_vis = fviz_silhouette(sil_km) + labs(title = 'K-Means Silhouette Plot')
```

```
##    cluster size ave.sil.width
## 1        1   17          0.12
## 2        2    5          0.19
## 3        3   28          0.22
## 4        4   17          0.12
## 5        5    2         -0.01
## 6        6    3          0.22
## 7        7    7          0.00
## 8        8    4          0.36
## 9        9   17          0.21
## 10      10    5          0.17
## 11      11    7          0.33
## 12      12   16          0.15
## 13      13   12          0.21
## 14      14   25          0.18
## 15      15    7          0.30
```

```
sing_vis = fviz_silhouette(sil_sing) + labs(title = 'Ward Linkage Silhouette Plot')
```

```
##    cluster size ave.sil.width
## 1        1    4         -0.07
## 2        2   11          0.45
## 3        3    7          0.19
## 4        4   14         -0.04
## 5        5   20          0.07
## 6        6    3          0.30
## 7        7    9          0.21
## 8        8   11          0.10
## 9        9   15          0.18
## 10      10    5          0.26
## 11      11   33          0.23
## 12      12   16          0.10
## 13      13    2         -0.01
## 14      14   15          0.25
## 15      15    7          0.29
```

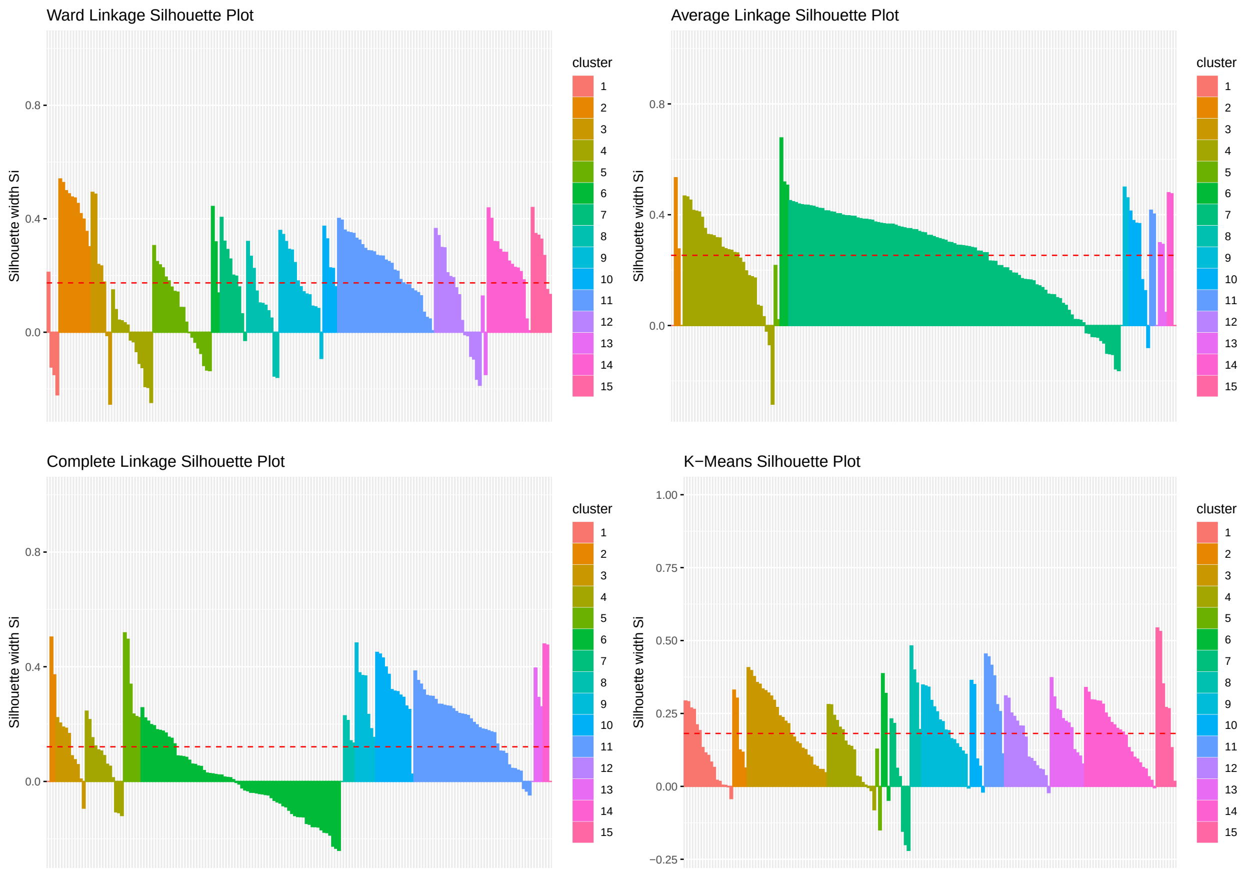
```
avg_vis = fviz_silhouette(sil_avg) + labs(title = 'Average Linkage Silhouette Plot')
```

```
##    cluster size ave.sil.width
## 1        1    1          0.00
## 2        2    2          0.40
## 3        3    1          0.00
## 4        4   31          0.25
## 5        5    2          0.12
## 6        6    3          0.57
## 7        7  113          0.25
## 8        8    1          0.00
## 9        9    2          0.48
## 10      10    7          0.25
## 11      11    2          0.41
## 12      12    1          0.00
## 13      13    3          0.21
## 14      14    2          0.48
## 15      15    1          0.00
```

```
cmplt_vis = fviz_silhouette(sil_cmplt) + labs(title = 'Complete Linkage Silhouette Plot')
```

```
##    cluster size ave.sil.width
## 1        1    1          0.00
## 2        2    2          0.44
## 3         3   10           0.11
## 4         4   13           0.06
## 5         5    6           0.34
## 6         6   68          -0.01
## 7         7    1           0.00
## 8         8    4           0.18
## 9         9    7           0.31
## 10       10   13           0.32
## 11       11   40           0.19
## 12       12    1           0.00
## 13       13    3           0.32
## 14       14    2           0.48
## 15       15    1           0.00
```

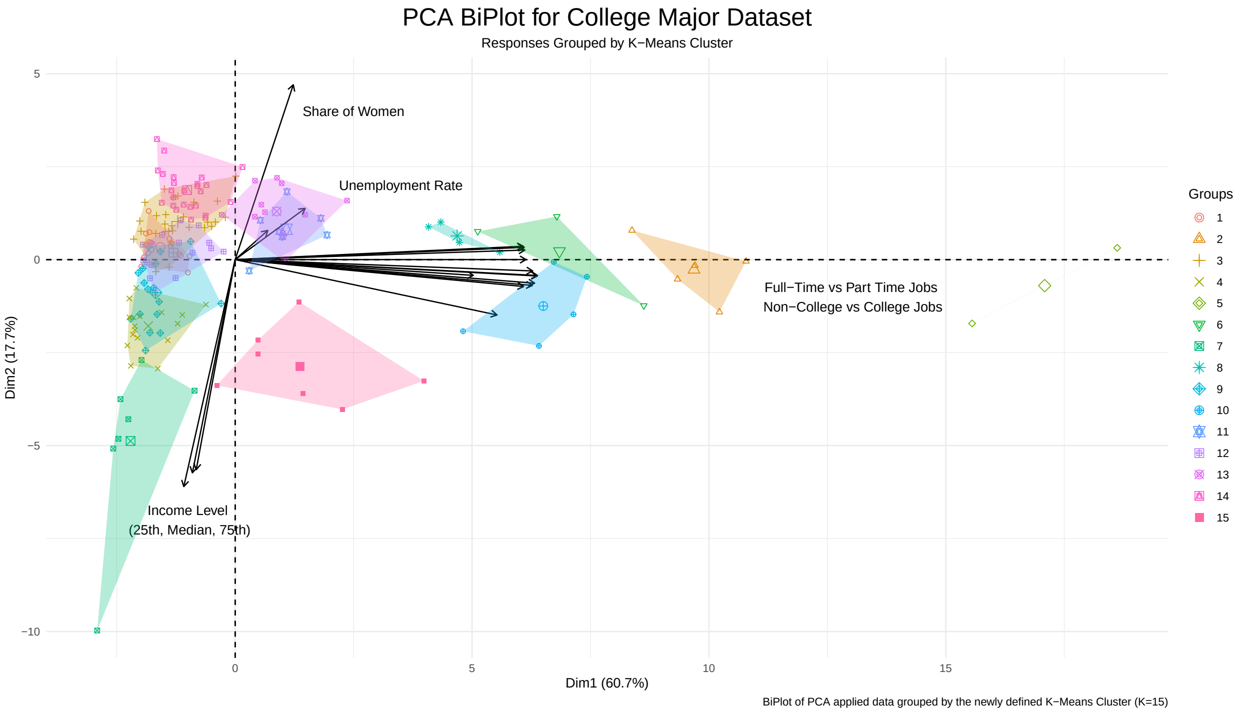
```
grid.arrange(sing_vis,
             avg_vis,
             cmplt_vis,
             km_vis,
             nrow = 2,
             ncol = 2)
```



Evaluation of the Data

This is the final biplot used for analysis - same algorithms as before just with greatest detail.

```
biplot_cluster = fviz_pca_biplot(prcomp(X),
                                col.var = 'black',
                                # var.scale = 0,
                                # alpha = 0.5,
                                label = FALSE,
                                habillage=(km_clust1$cluster))+
                                # addEllipses = TRUE)+
                  # scale_color_brewer(palette="Dark2")+
                  theme_minimal()+
                  annotate('text', label = "Share of Women", x = 2.5, y = 4)+
                  annotate('text', label = "Income Level\n (25th, Median, 75th)",
                           x = -1, y = -7)+
                  annotate('text',
                           label = "Full-Time vs Part Time Jobs\n Non-College vs College Jobs", x = 13, y = -1)+
                  annotate('text',
                           label = "Unemployment Rate", x = 3.5, y = 2)+
                  labs(title = "PCA BiPlot for College Major Dataset",
                       subtitle = 'Responses Grouped by K-Means Cluster',
                       caption = "BiPlot of PCA applied data grouped by the newly defined K-Means Cluster (K=15)") +
                  theme(plot.title = element_text(hjust = 0.5, size = 20),
                        plot.subtitle = element_text(hjust = 0.5))+
                  geom_convexhull(alpha = 0.3,
                                  aes(x = plot_dta_km1$PC1,
                                                  y = plot_dta_km1$PC2, fill = plot_dta_km1$Cluster),
                                  show.legend = FALSE)

biplot_cluster
```



Visualising the Cluster set based upon the differing types

This is some data preparation for the final table… alot of tidyverse operations I'm sure there was a better way to go about it but I used some brute force.

```r
cluster_category = dta_factored %>%
  select(Major_category, Median, Unemployment_rate) %>%
  mutate("Cluster" = as.factor(km_clust1$cluster),
         "Unemployment_rate" = round(Unemployment_rate*100,2))%>%
  group_by(Cluster) %>%
  summarise("Income" = scales::dollar(mean(Median)))%>%
            # "Unemployment" = paste(round(mean(Unemployment_rate),2),"%", sep = ""))%>%
  pivot_wider(names_from = Cluster,
              values_from = c(Income))
  # arrange(Cluster)


cluster_rate = dta_factored %>%
  select(Unemployment_rate) %>%
  mutate("Cluster" = as.factor(km_clust1$cluster),
         "Unemployment_rate" = round(Unemployment_rate*100,2))%>%
  arrange(Cluster)%>%
  group_by(Cluster)%>%
  pivot_wider(names_from = Cluster,
              values_from = Unemployment_rate,
              values_fn = mean)
colnames(dta_factored)
```

```
##  [1] "Rank"                 "Major_code"           "Major"
##  [4] "Total"                "Men"                  "Women"
##  [7] "Major_category"       "ShareWomen"           "Sample_size"
## [10] "Employed"             "Full_time"            "Part_time"
## [13] "Full_time_year_round" "Unemployed"           "Unemployment_rate"
## [16] "Median"               "P25th"                "P75th"
## [19] "College_jobs"         "Non_college_jobs"     "Low_wage_jobs"
## [22] "Income"
```

```r
cluster_women = dta_factored %>%
  select(ShareWomen) %>%
  mutate("Cluster" = as.factor(km_clust1$cluster),
         "ShareWomen" = ShareWomen*100)%>%
  arrange(Cluster)%>%
  group_by(Cluster)%>%
  pivot_wider(names_from = Cluster,
              values_from = ShareWomen,
              values_fn = mean)

cluster_employment = dta_factored %>%
  select(Employed) %>%
  mutate("Cluster" = as.factor(km_clust1$cluster))%>%
  arrange(Cluster)%>%
  group_by(Cluster)%>%
  pivot_wider(names_from = Cluster,
              values_from = Employed,
              values_fn = sum)



clustered_info = as.data.frame(rbind(cluster_category,
                                     paste(round(cluster_rate,2),"%", sep = ""),
                                     paste(round(cluster_women,2), "%", sep = ""),
                                     cluster_employment))
rownames(clustered_info) = c("Income", "Unemployment Rate", "Mean Share of Women",
                             "Total Employed")

#The final cluster table

pander::pander(clustered_info)
```

Table 2: Table continues below

	1	2	3	4
Income	$39,252.94	$35,680	$33,203.57	$51,470.59
Unemployment Rate	5%	7.35%	4.57%	3.48%
Mean Share of Women	52.25%	60.5%	61.59%	25.03%
Total Employed	105572	880153	365008	119361

Table 3: Table continues below

	5	6	7	8
Income	$34,750	$38,000	$74,285.71	$33,500
Unemployment Rate	7.8%	4.96%	7.77%	8.22%
Mean Share of Women	62.66%	87.75%	20.15%	59.01%
Total Employed	584167	448483	34485	384928

Table 4: Table continues below

	9	10	11	12
Income	$44,664.71	$42,400	$34,142.86	$41,343.75
Unemployment Rate	8.96%	8.26%	8.02%	10.03%
Mean Share of Women	27.32%	43.63%	67.81%	57.35%
Total Employed	125571	674187	324968	150724

	13	14	15
Income	$31,641.67	$30,792	$54,857.14
Unemployment Rate	7.54%	8.01%	6.04%
Mean Share of Women	59.32%	76.9%	23.8%
Total Employed	483231	293308	419053