Exploring The Impact of Education on Individuals - Preliminary Results

Steven Cao, Jesus Pacheco, Sharon Shen, Yaoxi Shi

Fall 2019

Part 1: Overview of the Project and Variables of Interest

1. Overview of the project

Current opinion holds strongly that education is key to realising the potential and growth of individuals, the association between increasing education and growth of economic capacity at both national and individual levels has been repeatedly noted (Psacharopoulos 1994), however, interpretations of why education leads to better outcomes are far from unanimous (Pritchett 2006).The main reason is that it is not clear what exactly about education leads to better outcomes, more granularity could be provided here as to the exact relationships between education and economic outcomes. We seek to explore such relationships through exploratory data analysis of the General Social Survey (GSS).We will be exploring variables for which we think there could be an intuitive explanation, e.g. comparing years of education and degree earned against income, in order to observe clusters which may or may not validate our intuitions. The goal, ultimately, will be to parse which aspects of education are most linked to economic outcome, with one potential and direct benefit being a better-informed policy towards education. In this preliminary report, we used variables from GSS database to investigate the relationship between education and economic outcomes as a sanity check of the relationship that has been stated in the literatures.

2. Considerations of the variables and the respective data

In the preliminary analysis, we chose three variables collected from 2016-2018 from General Social Survey (GSS) dataset: Education (in years), Income (in USD, adjusted for inflation), and Occupational Prestige (rated from 0-80, using the 2010 Census occupation classification). We chose these as our preliminary variables to perform analysis on for the reason that these variable was different enough to really cover a broad range of "ground" concerning "education and its effects, writ large", thus allow us to check the overall relationship between education and economic outcomes, also, the number of variables were manageable for our preliminary test.

However, a note: there were many missing values in the dataset which we've had to drop (around at least 2200, cf. variable summaries above; for comparison, the sample size of the dataset is 5215). We trust that this dropping of values does not impact the integrity (i.e. the representativeness) of our data, but we also accept that this is an assumption we are making, one which may infringe on the validity of our findings. In terms of an overall gloss of the variables, none seem unexpected.

- **Income** has a median to the left of the mean (37.5k versus 47.04k) suggesting a rightward skew (i.e. towards the upper bracket), which is expected.
- **Education** has a similar median and mean value, situated at ~14 years (for reference, 12 years is the amount to complete high school), which suggests that the majority of the sample have spent some time enrolled in higher education.
- **Occupational Prestige** is a score ranging from 0-80, taken from the 2010 Census Occupation Classification; a central measure of tendency isn't too interpretable here, but we can still point out that the mean and median are close to the exact middle point of that score range.

Here is a summary of the variables:

```
## [1] "   Income in USD"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     500   21250   37500   49538   67500  170000
```

```
## [1] "   Years of Education"
```

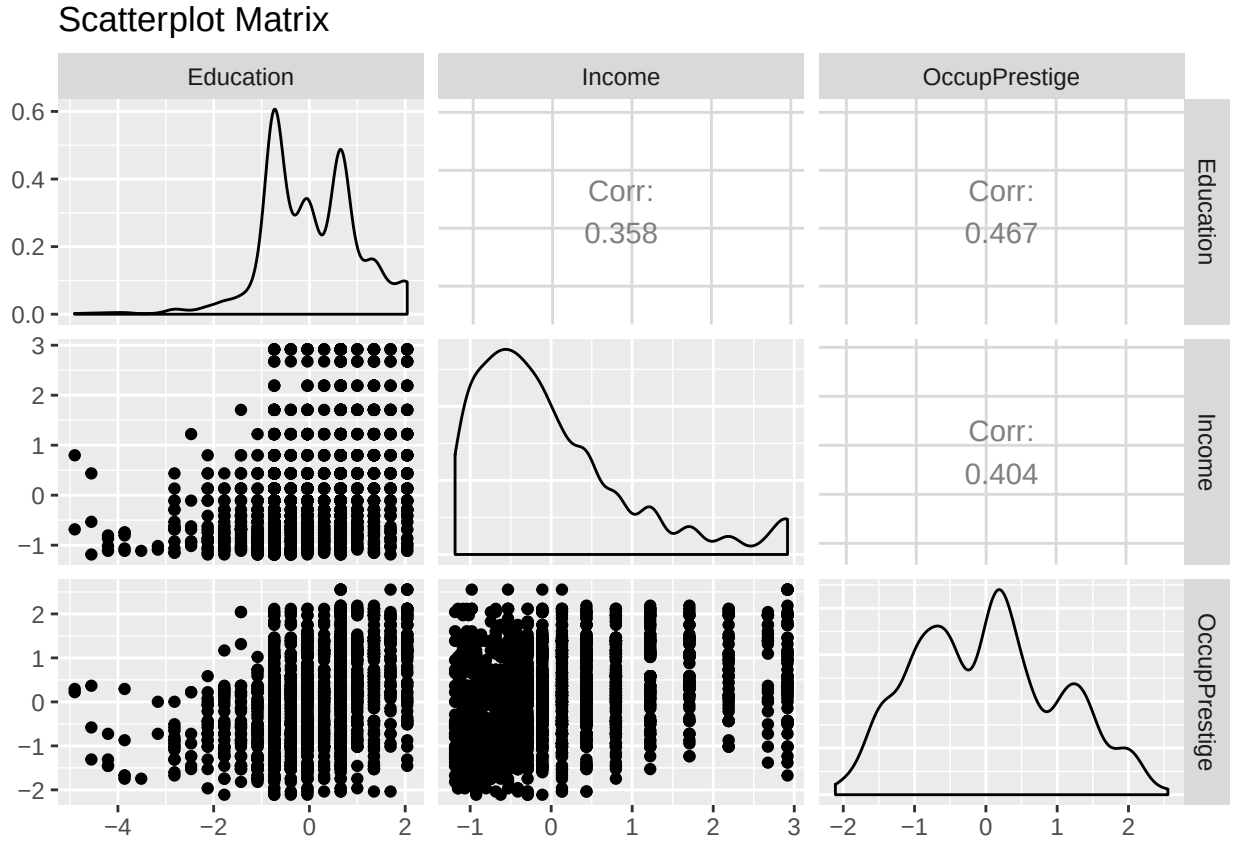
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    12.0    14.0    14.2    16.0    20.0
```

```
## [1] "   Occupational Prestige Score"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    16.0    35.0    45.0    45.2    54.0    80.0
```

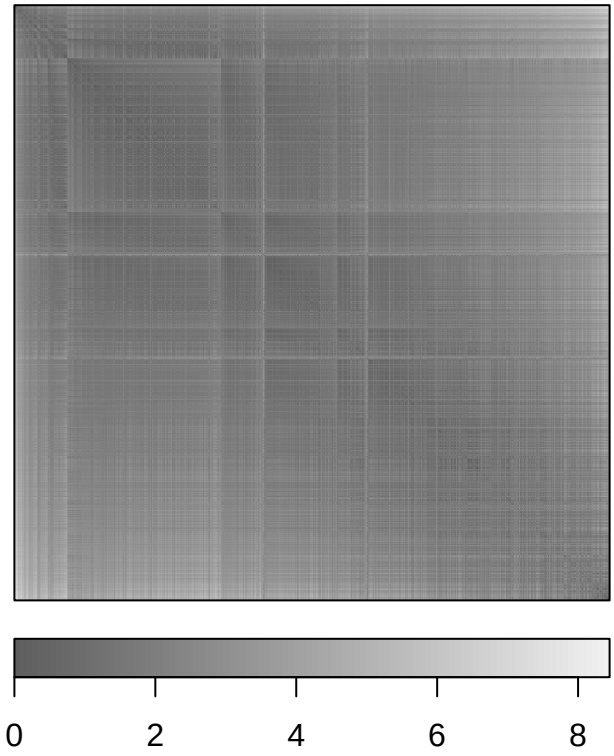
Part 2: Diagnosing Clusterability

(1) Informal Visualization & EDA



(2) ODI Plot

```
gss_scaled_1<-gss_scaled[1:3]
dist_mat<- dist(gss_scaled_1, method = "euclidean")
dissplot(dist_mat)
```



Comments on clusterability

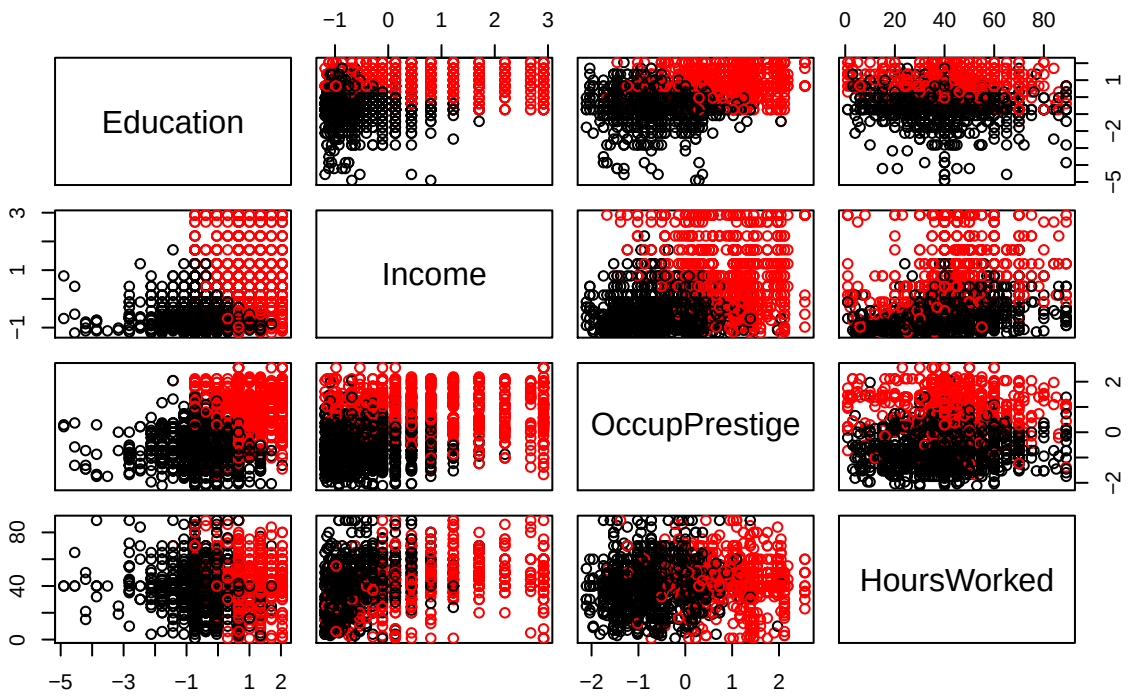
In a nutshell, the ODI plot shows an ambiguous block pattern, the color of the upper left and lower right corner are darker, though the boundaries are not sharp enough to distinguish each of the block, the pattern suggests that the data is not randomly distributed and could be clusterable in further anakysis.

Part 3: Partitioning Methods

(1) K-Means

```
km <- kmeans(gss_scaled_1,
             centers = 2,
             nstart = 10)
gss_scaled_1$clust <- as.factor(km$cluster)

pairs(~Education + Income + OccupPrestige + HoursWorked, data=gss_scaled_1, col=as.factor(gss_scaled_1$
      main="Scatterplot Matrix")
```



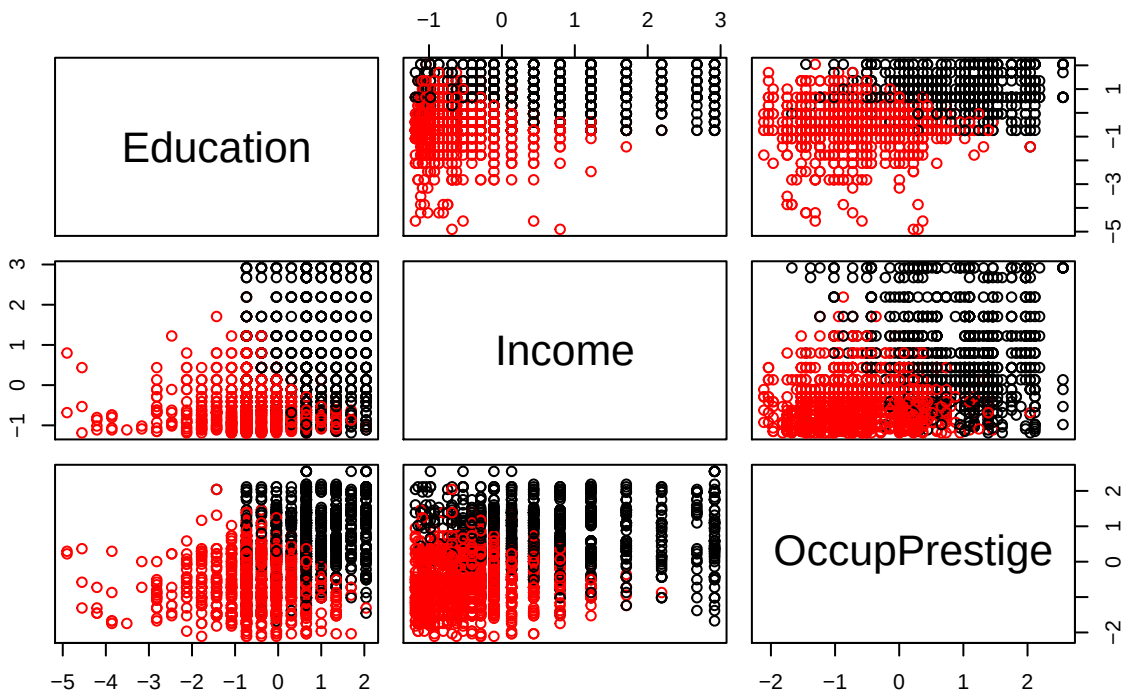

```
gss_scaled_1 %>% group_by(clust) %>% summarise(mean_educ = mean(Education), mean_income = mean(Income),
                                               mean_prest = mean (OccupPrestige))
```

```
## # A tibble: 2 x 4
##   clust mean_educ mean_income mean_prest
##   <fct>     <dbl>       <dbl>      <dbl>
## 1 1        -0.495      -0.467     -0.542
## 2 2         0.816       0.719      0.875
```

(2) PAM

```
pam <- pam(gss_scaled_1,
           k = 2)
gss_scaled_1$clust_pam <- as.factor(pam$clust)

pairs(~Education + Income + OccupPrestige, data=gss_scaled_1, col=as.factor(gss_scaled_1$clust_pam),
      main="Scatterplot Matrix")
```



(3) GMM

```
set.seed(123)
gmm <- mvnormalmixEM(as.matrix(gss_scaled_1[,1:3]),
            k = 2)
```

```
## number of iterations= 114
```

```
str(gmm)
```

```
## List of 9
##  $ x         : num [1:2548, 1:3] 0.6535 -0.7352 -0.7352 1.3479 -0.0409 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:3] "Education" "Income" "OccupPrestige"
##  $ lambda    : num [1:2] 0.649 0.351
##  $ mu        :List of 2
##   ..$ : num [1:3] -0.296 -0.476 -0.344
##   ..$ : num [1:3] 0.608 0.881 0.684
##  $ sigma     :List of 2
##   ..$ : num [1:3, 1:3] 0.8571 0.0641 0.1996 0.0641 0.2024 ...
##   ..$ : num [1:3, 1:3] 0.6972 0.0795 0.3491 0.0795 1.1814 ...
##  $ loglik    : num -9666
##  $ posterior : num [1:2548, 1:2] 0.0000000000114 0.8604773769476 0.9776685709337 0.9667214653028 0.0
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:2] "comp.1" "comp.2"
```

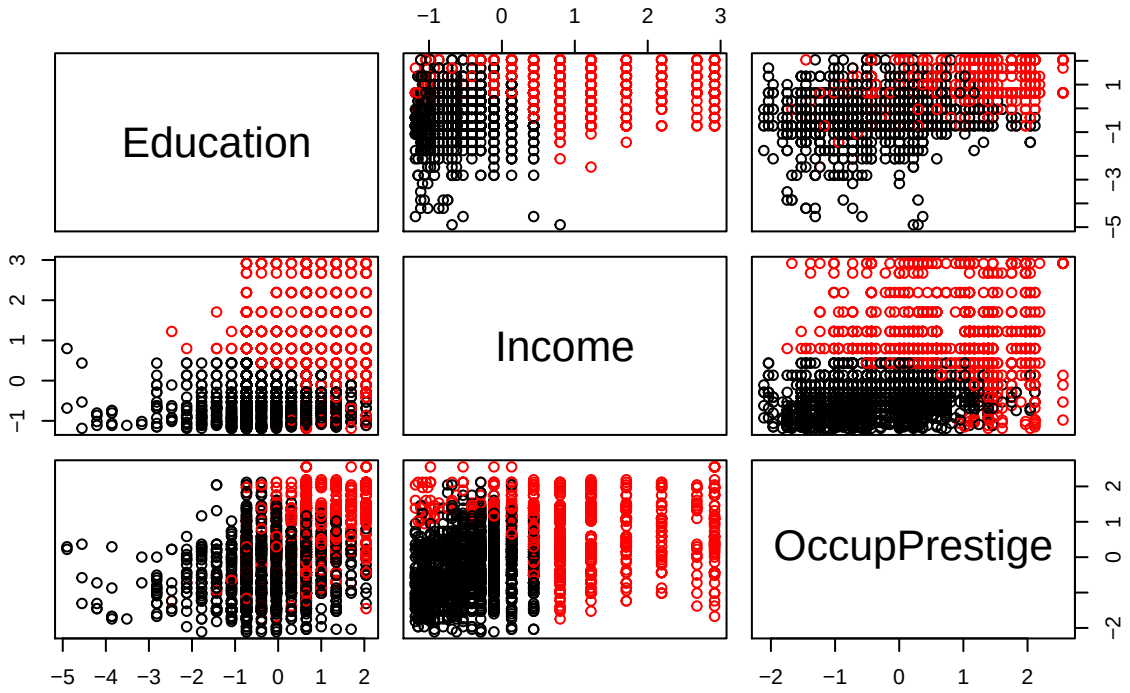
```
##  $ all.loglik: num [1:115] -19374 -10124 -10101 -10052 -9940 ...
##  $ restarts  : num 0
##  $ ft        : chr "mvnormalmixEM"
##  - attr(*, "class")= chr "mixEM"
```

```
gmm$cluster <- as.factor(ifelse(as.data.frame(gmm$posterior)$comp.1 > as.data.frame(gmm$posterior)$comp
summary(gmm$cluster)
```

```
##    1    2
## 1733  815
```

```
gss_scaled_1$clust2 <- as.factor(gmm$cluster)

pairs(~Education + Income + OccupPrestige, data=gss_scaled_1, col=as.factor(gss_scaled_1$clust2),
      main="Scatterplot Matrix")
```




```
gss_scaled_1 %>% group_by(clust2) %>% summarise(mean_educ = mean(Education), mean_income = mean(Income)
                                                mean_prest = mean (OccupPrestige))
```

```
## # A tibble: 2 x 4
##   clust2 mean_educ mean_income mean_prest
##   <fct>      <dbl>       <dbl>      <dbl>
## 1 1         -0.322      -0.487     -0.367
## 2 2          0.752       1.04       0.835
```

Internal Validation

```
set.seed(123)
samp <- sample_n(gss_scaled_1[1:100,1:3],100)
internal <- clValid(samp, nClust = 2:10,
                    clMethods = c("kmeans", "pam", "model"),
                    validation = "internal")
```

```
## Warning in clValid(samp, nClust = 2:10, clMethods = c("kmeans", "pam",
## "model"), : rownames for data not specified, using 1:nrow(data)
```

```
summary(internal)
```

```
##
## Clustering Methods:
##  kmeans pam model
##
## Cluster sizes:
##  2 3 4 5 6 7 8 9 10
##
## Validation Measures:
##                            2       3       4       5       6       7       8       9      10
##
## kmeans Connectivity   16.461  31.068  45.687  49.865  56.002  59.817  59.040  66.253  78.271
##        Dunn            0.092   0.086   0.057   0.063   0.115   0.129   0.172   0.078   0.078
##        Silhouette      0.345   0.325   0.265   0.272   0.281   0.290   0.291   0.288   0.269
## pam    Connectivity   15.893  32.875  62.298  51.567  57.844  66.411  70.340  72.662  82.071
##        Dunn            0.092   0.073   0.026   0.080   0.016   0.048   0.080   0.119   0.096
##        Silhouette      0.344   0.315   0.241   0.270   0.276   0.248   0.279   0.246   0.265
## model  Connectivity   20.429  31.813  42.950  40.486  94.618  66.712 141.348 131.922  79.257
##        Dunn            0.034   0.039   0.044   0.140   0.022   0.095   0.072   0.043   0.089
##        Silhouette      0.341   0.321   0.283   0.290   0.104   0.266  -0.012   0.011   0.275
##
## Optimal Scores:
##
##              Score  Method Clusters
## Connectivity 15.893 pam    2
## Dunn          0.172 kmeans 8
## Silhouette    0.345 kmeans 2
```

Part 4: Comments on analyses

According to the ODI diagram, we set K=2 for partitioning, and we used K-means, PAM and GMM algorithms for our analysis. From the results shown above, K-Means and PAM both yielded near-identical results, suggesting that the use of both would be redundant. Considering that PAM is good in handling outliers (and that the scaled data does not have outliers) and other heavy skewing factors in the data, this outcome was about expected.

One of the main observed differences between GMM and K-Means is the dispersion of data classified. In most cases, GMM allows for observations belonging to different clusters to be interspersed among one another more generously than K-Means (e.g. the scatterplot comparing Education and Occupational Prestige). We believe that overall, GMM should fit our exploratory data analysis better than K-Means for the simple reason that there is bound to be a lot of overlap within the data (again, there is no clear boundaries for clustering among any of the preliminary variables).

Also, for all our partioning methods, the internal validation strongly suggested that k=2 would yield the best

clustering results.

As for the interpretation of the clustering results, we find that the data points in one cluser have obviously higher value in years of education, income, and occupation prestige than another cluster, suggesting thet people in the higher income brackets will tend to also have higher years of education and also have ocupations with higher prestige scores. But we haven't explore more granularity of the relationship. For the next step, we are going to include more variables related to education such as major, degree earned etc to investigate what specific aspects of education are more related to one's economic outcome.