MAC40800 - Project

Exploring the Impact of Education on Individuals

Cao, Pacheco, Shen, & Shi

Why are STEM majors so popular now?

Will taking UML lead you to a better career?

What are the prospects out of getting a PhD?

Introduction

Background

Explore the paths people take in obtaining various socio-economic advantages

+

Literature Review

Individual features such as education attainment could be used to explain income level

→

Our Approach

Leverage data from General Social Survey 2016-18 to uncover patterns and themes among respondents

GSS at a glance…

Search Data

Keyword All

education

Filter by Years: 2000 - 2018

or Select Specific Years

Filter by Module / Subject:

Choose Modules / Subjects

SEARCH

Clear all filters

Subjects matching 'education'

Education

45 Results Matching Criteria

Show results that match on:

All | Variable Name | Variable Description | Survey Question

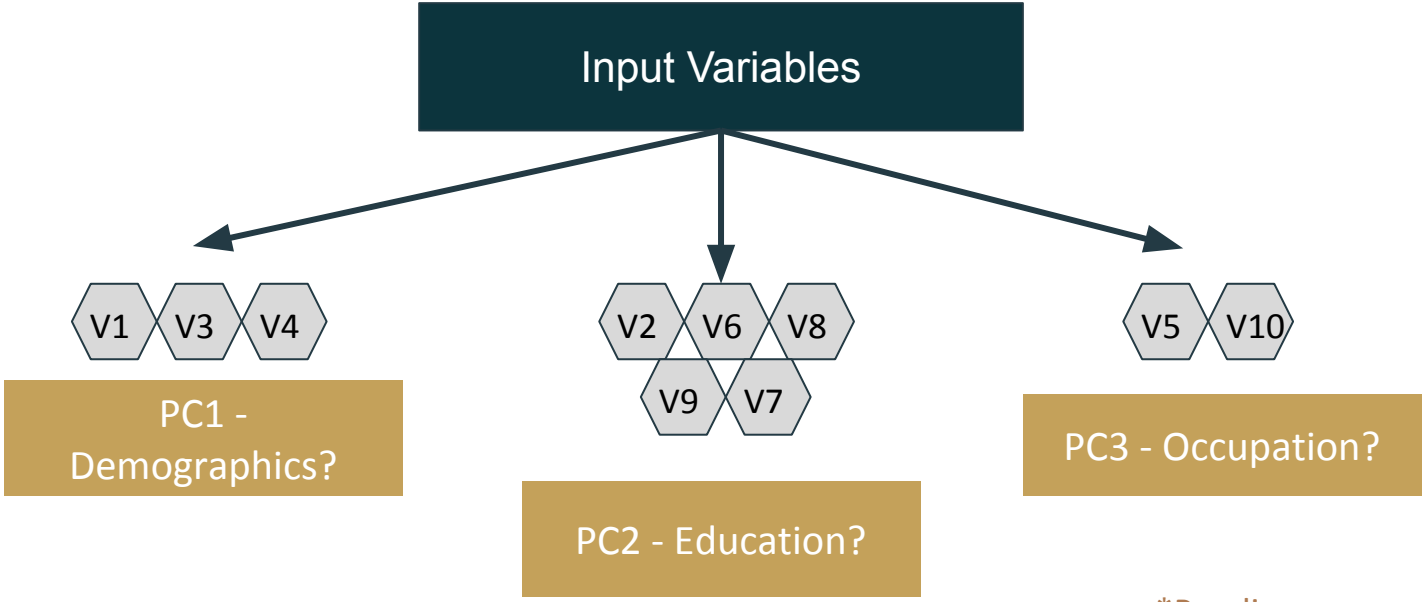
Save Search | Collapsed View | Expanded View

+ ALL	Variable Name	Description	Years Available 1972 - 2018
✓	voedcol	Non-college postsecondary **education** (voednme1)	
✓	voedncol	Non-college postsecondary **education** (voednme2)	
✓	nateduc	Improving nations **education** system	

Feature Space

- Education
  - Highest year of school completed
  - Rs highest degree
  - The field of degree r earned
  - Type of college respondent attended
  - College major
  - R has taken any college-level sci course
- Income
  - Respondents income
  - Rs federal income tax
  - Opinion of family income
- Demographic
  - Respondents sex
  - R's socioeconomic index
  - Age of respondent
  - Race of respondent
- Occupation
  - Respondent’s occupation prestige score
  - Labor force status
  - Respondent’s NAICS industry code
  - Numbers of hours work per week
  - R self-emp or works for somebody
  - Govt or private employee

Methodology - Step I: PCA*



*Pending process

Methodology - Step II: EDA

Initial Variable Selection:

- Education
- Income
- Occupation Prestige
- Hours worked

Diagnose Clusterability

- ODI: *lack of shaded diagonal blocks*
- Hopkins Statistics: *H=0.17*

Clustering

- K-means (2-cluster model)
- GMM (2-cluster model)

Results - Data wrangling

Respondent's Income

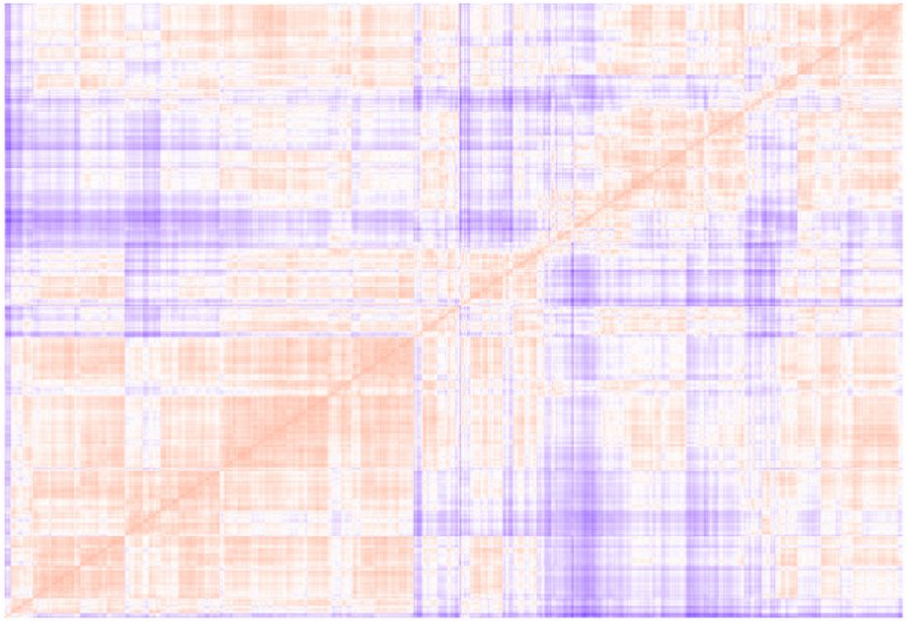
under $1 000	$1 000 to 2 999	$3 000 to 3 999	$4 000 to 4 999	$5 000 to 5 999	$6 000 to 6 999
32	49	41	34	36	32
$7 000 to 7 999	$8 000 to 9 999	$10000 to 12499	$12500 to 14999	$15000 to 17499	$17500 to 19999
37	51	82	76	66	71
$20000 to 22499	$22500 to 24999	$25000 to 29999	$30000 to 34999	$35000 to 39999	$40000 to 49999
116	95	148	184	173	264
$50000 to 59999	$60000 to 74999	$75000 to $89999	$90000 to $109999	$110000 to $129999	$130000 to $149999
219	234	148	124	74	48
$150000 to $169999	$170000 or over	refused	dk	na	IAP
27	87	0	0	0	0

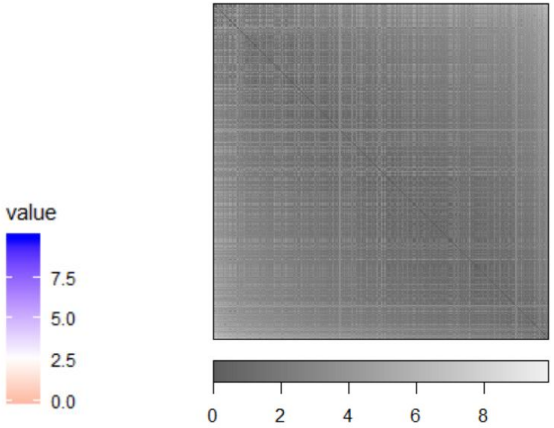
- Ordinal to continuous by taking the median of each income bracket

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
500	13750	37500	47789	67500	170000

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3918	0.4785	0.9430	1.0196	0.9570	5.8974

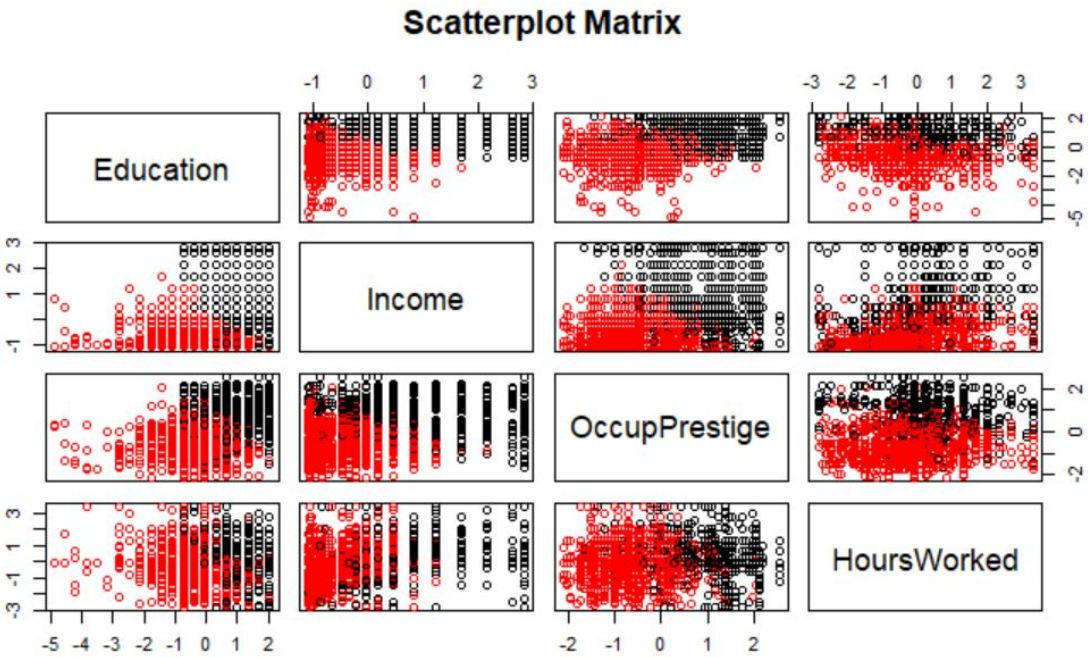
Results - ODI





- Based on 2548 Observations with 4 variables

Results - Clustering



- K-means with k=2
- There is a visible distinction between groups, especially concerning education and income
- A GMM model shows similar pattern

Next Steps

- Add more variables
- Perform PCA
- Reiterate clustering process
- Other suggestions?