# 情報可視化ソフトウェア**Hidden**によるデータ分析
## ～ 教育に対する社会的要因の関係～

**1.** 動機

一般に子供の学力について、親の年収や学歴、住んでいる地域など様々な要因が影響していると言われている。確かにそれらの要因は子供の学力に影響しているように感じるが、実際にはどの程度それぞれの要因が影響しているのか気になったので、このテーマを選んだ。

**2.** データについて

Rdatasets、AERパッケージのCollegeDistanceというcsvデータをダウンロードした。
アメリカ合衆国で1980年に教育省によって実施された「High School and Beyond Survey」からのクロスセクションデータで、1980年の調査の後、1986年にフォローアップが行われたもの。
この調査には、約1,100校の高校からの学生が含まれている。
生のデータは図1の通り。

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | rownames | gender | ethnicity | score | fcollege | mcollege | home | urban | unemp | wage | distance | tuition | education | income | region |
| 2 | 1 | male | other | 39.1500015 | yes | no | yes | yes | 6.19999981 | 8.09000015 | 0.2 | 0.88915002 | 12 | high | other |
| 3 | 2 | female | other | 48.8699989 | no | no | yes | yes | 6.19999981 | 8.09000015 | 0.2 | 0.88915002 | 12 | low | other |
| 4 | 3 | male | other | 48.7400017 | no | no | yes | yes | 6.19999981 | 8.09000015 | 0.2 | 0.88915002 | 12 | low | other |
| 5 | 4 | male | afam | 40.4000015 | no | no | yes | yes | 6.19999981 | 8.09000015 | 0.2 | 0.88915002 | 12 | low | other |
| 6 | 5 | female | other | 40.4799995 | no | no | no | yes | 5.5999999 | 8.09000015 | 0.40000001 | 0.88915002 | 13 | low | other |
| 7 | 6 | male | other | 54.7099991 | no | no | yes | yes | 5.5999999 | 8.09000015 | 0.40000001 | 0.88915002 | 12 | low | other |
| 8 | 7 | female | other | 56.0699997 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.40000001 | 0.84987998 | 13 | low | other |
| 9 | 8 | female | other | 54.8499985 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.40000001 | 0.84987998 | 15 | low | other |
| 10 | 9 | male | other | 64.7399979 | yes | no | yes | yes | 5.9000001 | 8.09000015 | 3 | 0.88915002 | 13 | low | other |
| 11 | 10 | female | other | 56.0600014 | no | no | yes | yes | 5.9000001 | 8.09000015 | 3 | 0.88915002 | 15 | low | other |
| 12 | 11 | female | other | 42.2200012 | no | no | yes | yes | 5.9000001 | 8.09000015 | 3 | 0.88915002 | 12 | high | other |
| 13 | 12 | female | afam | 61.1800003 | no | yes | yes | yes | 5.9000001 | 8.09000015 | 3 | 0.88915002 | 14 | high | other |
| 14 | 13 | male | other | 59.8499985 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 15 | low | other |
| 15 | 14 | female | other | 58.7700005 | yes | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 17 | high | other |
| 16 | 15 | female | afam | 53.7200012 | yes | yes | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 14 | low | other |
| 17 | 16 | male | other | 61.5200005 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 15 | low | other |
| 18 | 17 | female | other | 52.5299988 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 12 | high | other |
| 19 | 18 | female | other | 45.0099983 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 12 | low | other |
| 20 | 19 | female | other | 57.7099991 | no | no | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 16 | low | other |
| 21 | 20 | female | other | 59.3600006 | yes | yes | yes | no | 7.19999981 | 8.85000038 | 0.1 | 0.84987998 | 16 | high | other |

図1：サイトからダウンロードした生のデータ(csv)

数値でないデータが多く含まれていたため、図2のように編集した。(rownamesは削除)

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric |
| 2 | gender | ethnicity | score | fcollege | mcollege | home | urban | unemp | wage | distance | tuition | education | high_income | region_is_west |
| 3 | 1 | 2 | 39.1500015 | 1 | 0 | 1 | 1 | 6.19999981 | 8.09000015 | 0.2 | 0.88915002 | 12 | 1 | 0 |
| 4 | 0 | 2 | 42.2200012 | 0 | 0 | 1 | 1 | 5.9000001 | 8.09000015 | 3 | 0.88915002 | 12 | 1 | 0 |
| 5 | 0 | 2 | 50.9300003 | 0 | 0 | 1 | 0 | 7.69999981 | 7.03999996 | 0.5 | 0.903 | 13 | 0 | 0 |
| 6 | 0 | 2 | 68.5800018 | 1 | 1 | 1 | 1 | 5.09999991 | 8.85000038 | 0.5 | 0.84987998 | 16 | 1 | 0 |
| 7 | 1 | 2 | 70.0999985 | 1 | 0 | 1 | 0 | 6.19999981 | 8.09000015 | 0.60000002 | 0.88915002 | 17 | 1 | 0 |
| 8 | 1 | 1 | 49.6500015 | 0 | 0 | 0 | 0 | 7.19999981 | 7.03999996 | 0.1 | 0.903 | 15 | 0 | 0 |
| 9 | 1 | 2 | 57.7400017 | 0 | 0 | 1 | 0 | 5.5 | 8.09000015 | 1.5 | 0.88915002 | 16 | 1 | 0 |
| 10 | 0 | 2 | 60.9300003 | 0 | 0 | 1 | 0 | 6.5 | 8.09000015 | 1.5 | 0.88915002 | 14 | 0 | 0 |
| 11 | 1 | 2 | 65.3099976 | 1 | 0 | 1 | 0 | 7.5 | 8.09000015 | 0.5 | 0.88915002 | 16 | 0 | 0 |
| 12 | 0 | 2 | 45.1399994 | 0 | 0 | 1 | 0 | 5.9000001 | 7.09000015 | 1.20000005 | 1.38568008 | 12 | 1 | 0 |
| 13 | 0 | 2 | 40.2799988 | 0 | 0 | 1 | 0 | 6.09999991 | 7.69000006 | 2.5 | 1.11201 | 12 | 1 | 0 |
| 14 | 1 | 0 | 50.8100014 | 0 | 1 | 1 | 0 | 5.4000001 | 8.09000015 | 1.89999998 | 0.88915002 | 16 | 0 | 0 |
| 15 | 0 | 2 | 50.9500008 | 0 | 0 | 1 | 0 | 8.69999981 | 8.85000038 | 1.5 | 0.84987998 | 12 | 0 | 0 |
| 16 | 1 | 2 | 54.2599983 | 0 | 0 | 1 | 0 | 6.5 | 8.09000015 | 0.80000001 | 0.88915002 | 14 | 0 | 0 |
| 17 | 1 | 2 | 38.1899986 | 0 | 0 | 1 | 0 | 7.80000019 | 7.69000006 | 1.79999995 | 1.11201 | 12 | 0 | 0 |
| 18 | 1 | 2 | 52.9799995 | 1 | 1 | 1 | 0 | 7.19999981 | 8.85000038 | 0.80000001 | 0.84987998 | 13 | 1 | 0 |
| 19 | 1 | 2 | 66.4899979 | 0 | 0 | 1 | 0 | 4.4000001 | 8.09000015 | 1.5 | 0.88915002 | 17 | 1 | 0 |
| 20 | 1 | 2 | 52.6399994 | 0 | 1 | 1 | 0 | 5.09999991 | 8.85000038 | 0.1 | 0.84987998 | 15 | 1 | 0 |
| 21 | 0 | 2 | 57.1199989 | 0 | 0 | 1 | 0 | 10.3999996 | 7.69000006 | 0 | 1.11201 | 15 | 1 | 0 |

図2：加工後のデータ(.csv)

図2:加工後のデータ(.csv) の各パラメータは以下の通りである。

① **gender**
男性:1　女性:0

② **ethnicity**
アフリカ系アフリカ人:0　ヒスパニック:1　その他:2

③ **score**
高校3年生時のテストの結果

④ **fcollege**
父親が大学を卒業している:1　卒業していない:0

⑤ **mcollege**
母親が大学を卒業している:1　卒業していない:0

⑥ **home**
家族が自宅を所有している:1　所有していない:0

⑦ **urban**
学校が都市部にある:1　都市部にない:0

⑧ **unemp**
調査当時(1980年)のその州の失業率(%)

⑨ **wage**
調査当時(1980年)のその州の製造業の時給(ドル)

⑩ **distance**
4年制大学からの距離(10マイル単位)

⑪ **tuition**
その州の州立四年制大学の平均授業料(1000ドル単位)

⑫ **education**
調査時点までの教育年数。
高校卒業:12
専門学校:13
準学士(AA):14
学士(BA):16
大学院教育の一部を受けた人:17
大学院の学位を持つ修士:18

⑬ **high_income**
家族収入が年間2万5000ドルを超えている:1　超えていない:0

⑭ **region_is_west**
その地域が西部である:1　その他の地域である:0

また約5000行という非常に大きなデータで実行したところ、非常に重かった。全てのデータを使用する必要性はないと考え、約500行までデータを削減して解析を行った。

## 3. 結果

①初期状態

まずはデータを読み込んで、何もパラメータを変更しない状態で観察してみる。
結果は図3のようになった。

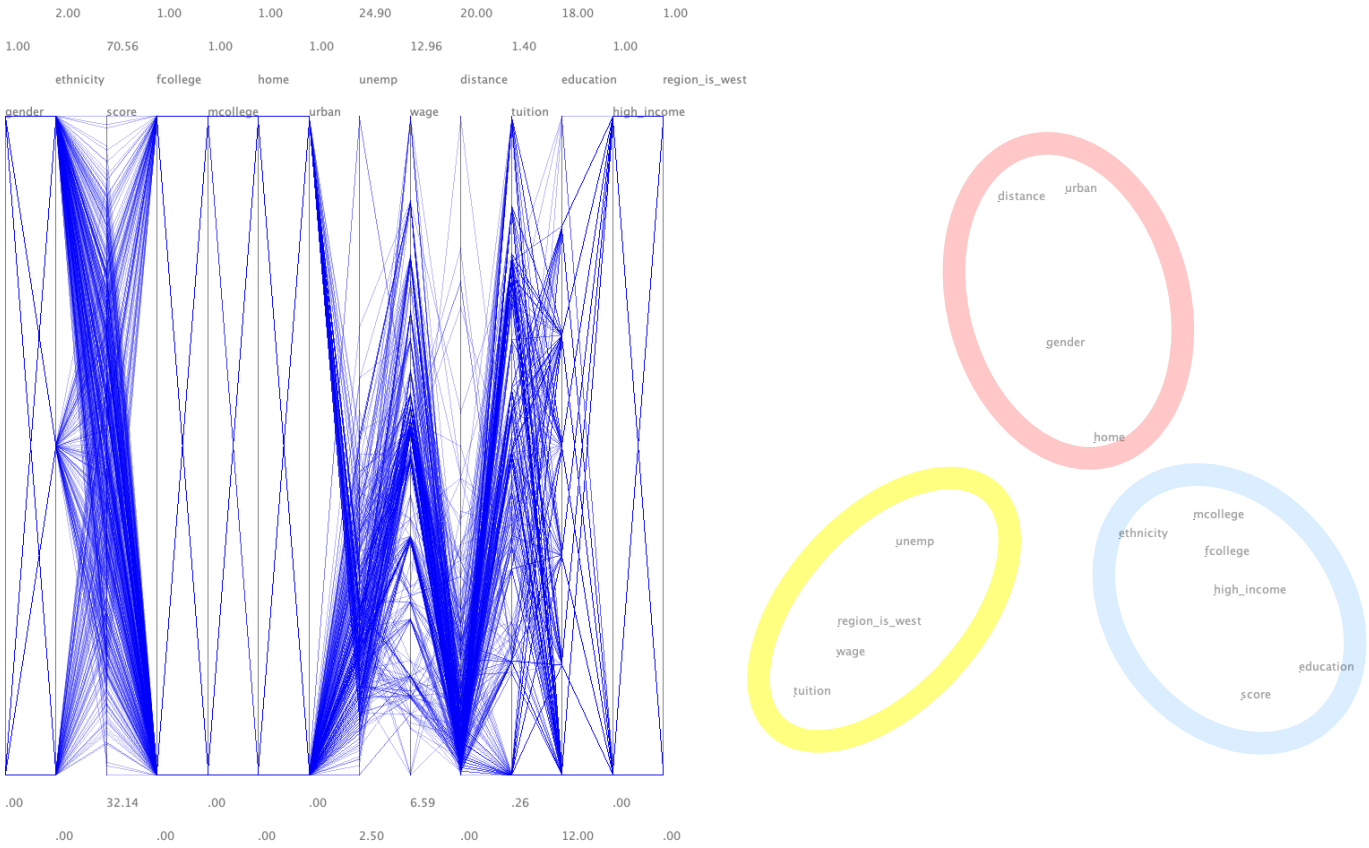


図3：解析結果(初期状態)

＜考察＞

右側の散布図に注目すると、これらのパラメータは大きく分けると**3つに分類**できそうだと感じた。

① **distance, urban, gender, home**

② **unemp, region_is_west, wage, tuition**

③ **ethnicity, mcollege, fcollege, high_income, education, score**

ただ、データ数が非常に多く、左側のPCPから分析を行うことは難しい。そのため、この後右側のスライダを動かして、相関の強いものを順に見ていくことにした。

また、今回は**14**の変数のうち半数以上が**2**値、またはカウントデータであり、今回のように可視化する際に**2**値データの軸間では複数の線分が大量に重なってしまっている場所があり、PCPからはあまりたくさんの情報を得られないのではないかと考えた。そのため、量的尺度の軸と**2**値の軸間で比較するなど、情報量の多い軸に特に注目してみていくことにした。

②右側のスライダーをあげる
右側のスライダーを**4**分の**1**ほど動かし、クラスタの数を**2**にした。
結果は図4のようになった。

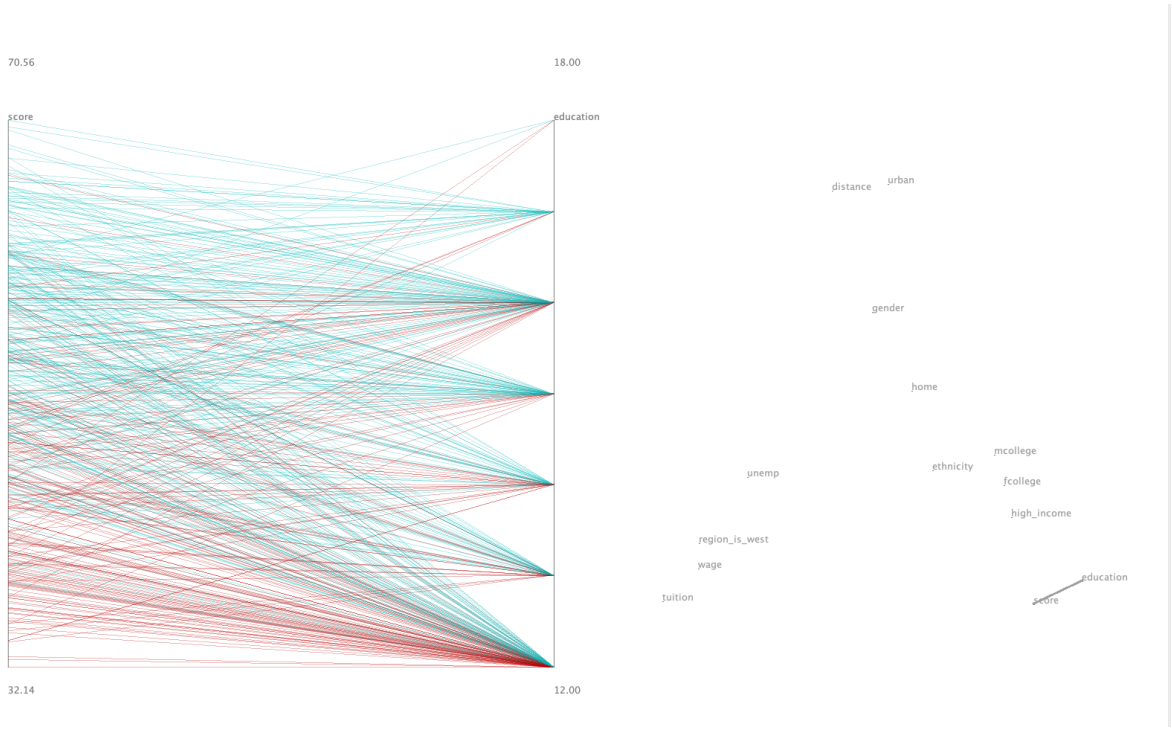


図4：解析結果(クラスタ数2)

＜考察＞
右側のスライダーをあげて一番最初に線分で結ばれたのは**education**と**score**だった。
クラスタリングの数を調節してみたところ、**2**の時に見やすいPCPの図が得られた(図4左側)。
Educationが上がるにつれて、青い線の割合が増加していることがわかる。
(educationが18のものは、母数が極端に少ないのでこの傾向はあてはまっていないが、データ数を十分に揃えればこれに当てはまるのではないかと考える。)
赤い線はスコアが下位のもの、青い線はスコアが上位のものであることがscore軸から分かり、**education**の数値が上がれば上がるほど青い線の割合が増加する、つまりスコアが高くなることがわかった。これにより、高校**3**年生時点での成績が高ければ高いほど、卒業後の教育年数が長くなることが分かる。

クラスタ数を4に変更したものが図5である。こちらもeducationが大きくなるにつれて、赤い線の割合が減少しており、同様の考察が得られる。

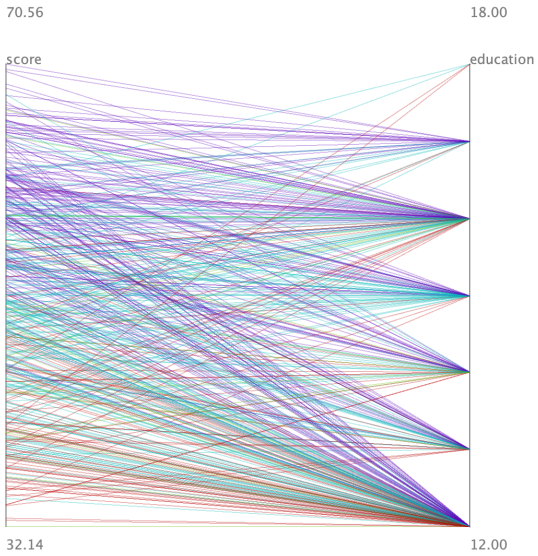


図5：クラスタ数を4にした場合の解析結果(score軸、education軸)

③右側のスライダーを上げて変化を見る
出力結果に変化があるまで右側のスライダーを動かしたところ、図6のような結果になった。

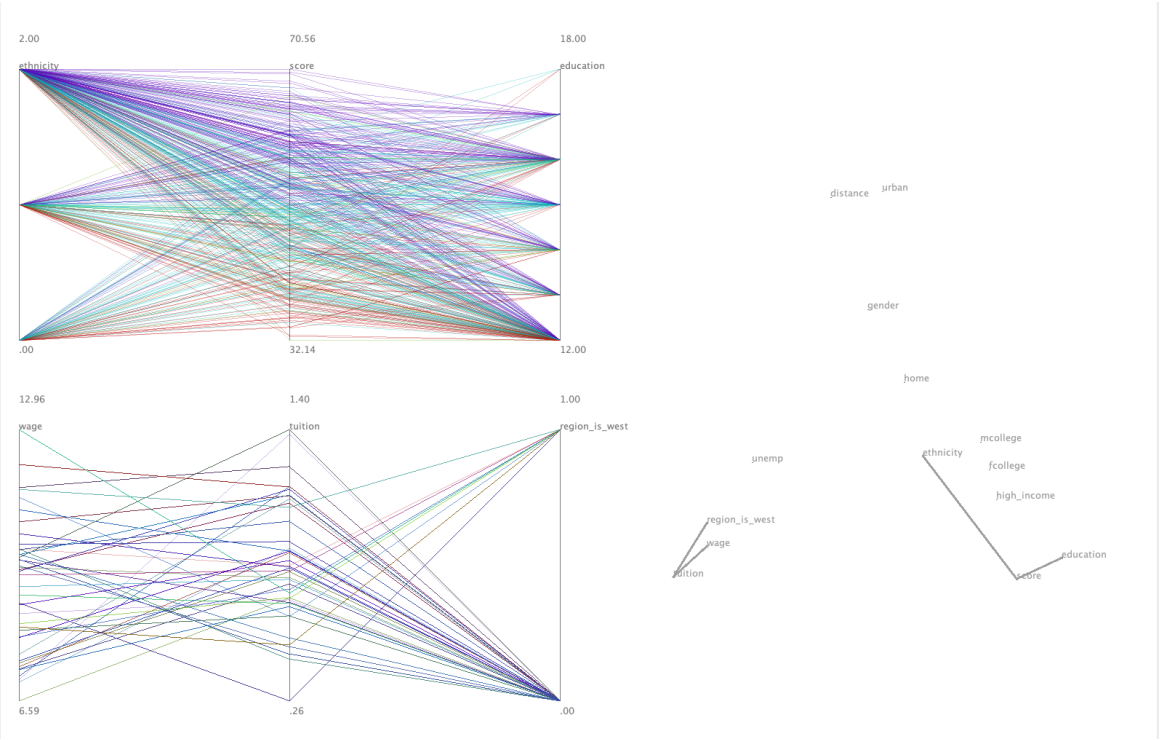


図6：解析結果(クラスタ数4)

＜考察＞
右側のスライダーを動かすと、次にPCPに軸として現れたのは**ethnicity**と**score**だった。
また、**wage**, **tuition**,**region_is_east** も線分で結ばれたが、値が同じものが多いため、データ数に対して引かれる線分の数が非常に少なく、線分が重なっているためクラスタ数を変化させてもあまり良い情報は得られないと考えたため、ここでは取り上げない。

ethnicityとscoreについて、図6の左側のPCPに注目する。
紫色の線分(=高得点群)の割合は以下の順に多くなっている。

その他 > ヒスパニック > アフリカ系アフリカ人

また、赤色の線分(=低得点群)の割合は以下の順に多くなっている。

アフリカ系アフリカ人 ≧ ヒスパニック > その他

このことから人種と高校時代の成績には密接な関係があると推測できる。
また、図6のようにクラスタ数を**2**に変更した場合でも、比較的分かりやすくクラスタリングできていることが確認できた。

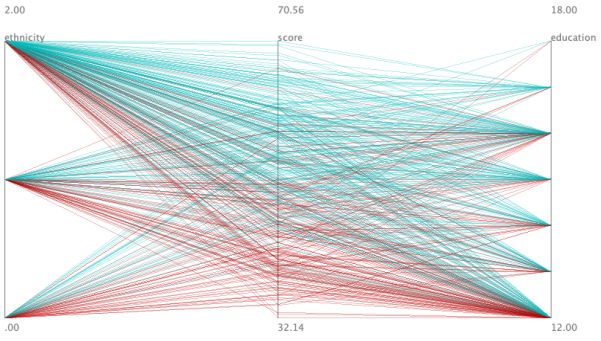


図6：クラスタ数2の場合のPCP

④さらに右側のスライダーを上げて変化を見る
③の状態からさらに右側のスライダーを動かして変化を見た。結果は図7のようになった。

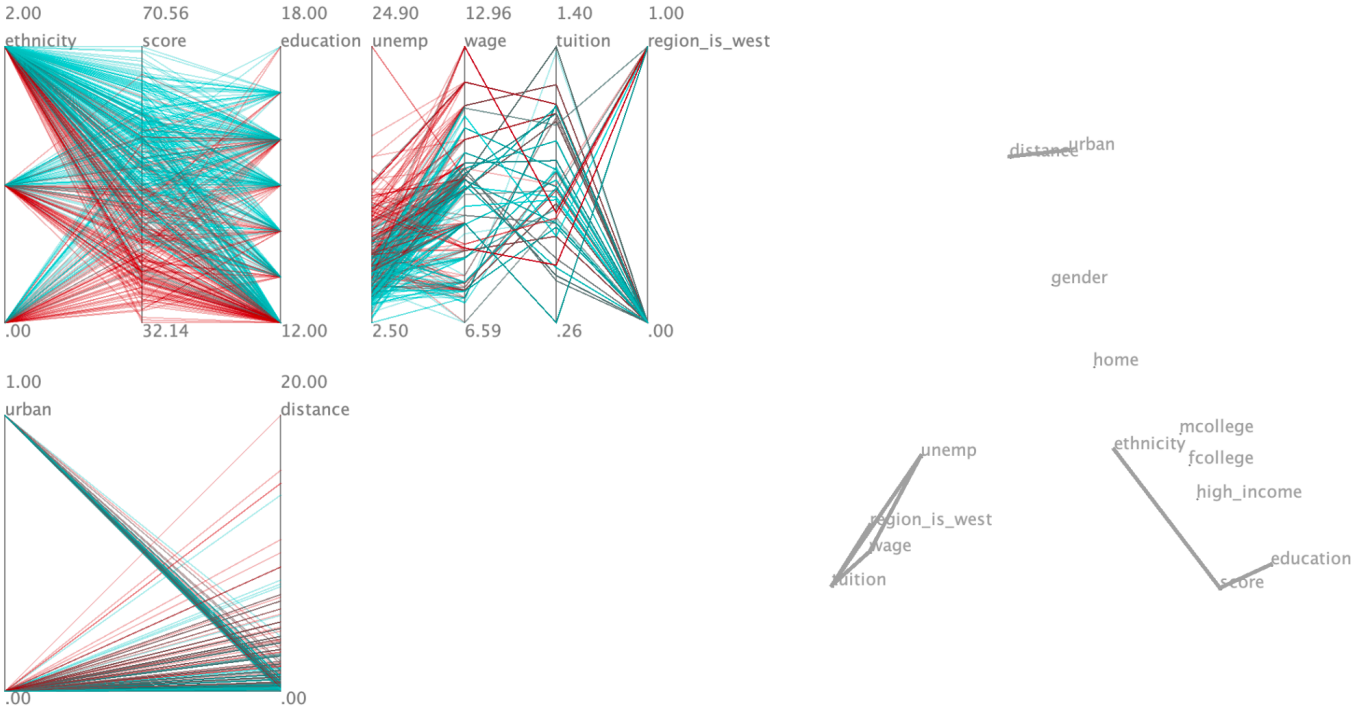


図7：解析結果(クラスタ数2)

＜考察＞
図7PCPの右上に注目する。クラスタ数を**2**設定し、**unemp**と**wage**の軸に注目すると、比較的綺麗にクラスタリングすることができている。これは、失業率が高いほど賃金が比較的高い、つまり失業率と賃金には正の相関関係があることが読み取れる。

また、左下のPCPに注目するためにクラスタ数を3にした結果が図8である。
これは**urban**と**distance**の関係性を示しており、都市部であるほど、州立大学までの距離が小さく、都市部でない場合は州立大学までの距離が極端に大きい場合が多いことが読み取れる。
このことから、アメリカでは都市部に大学が比較的たくさん集まっていることが考えられる。

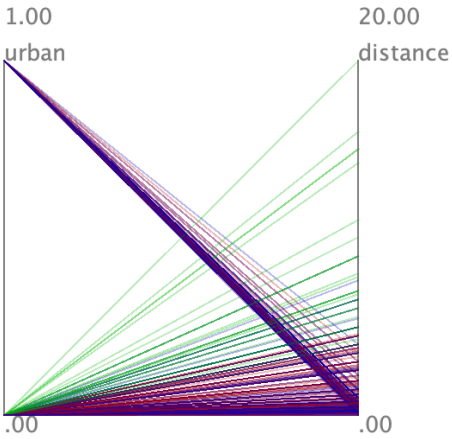


図8：解析結果(クラスタ数3)

⑤さらに右側のスライダーを上げて変化を見る
④の状態からさらに右側のスライダーを動かして変化を見た。結果は図9のようになった。

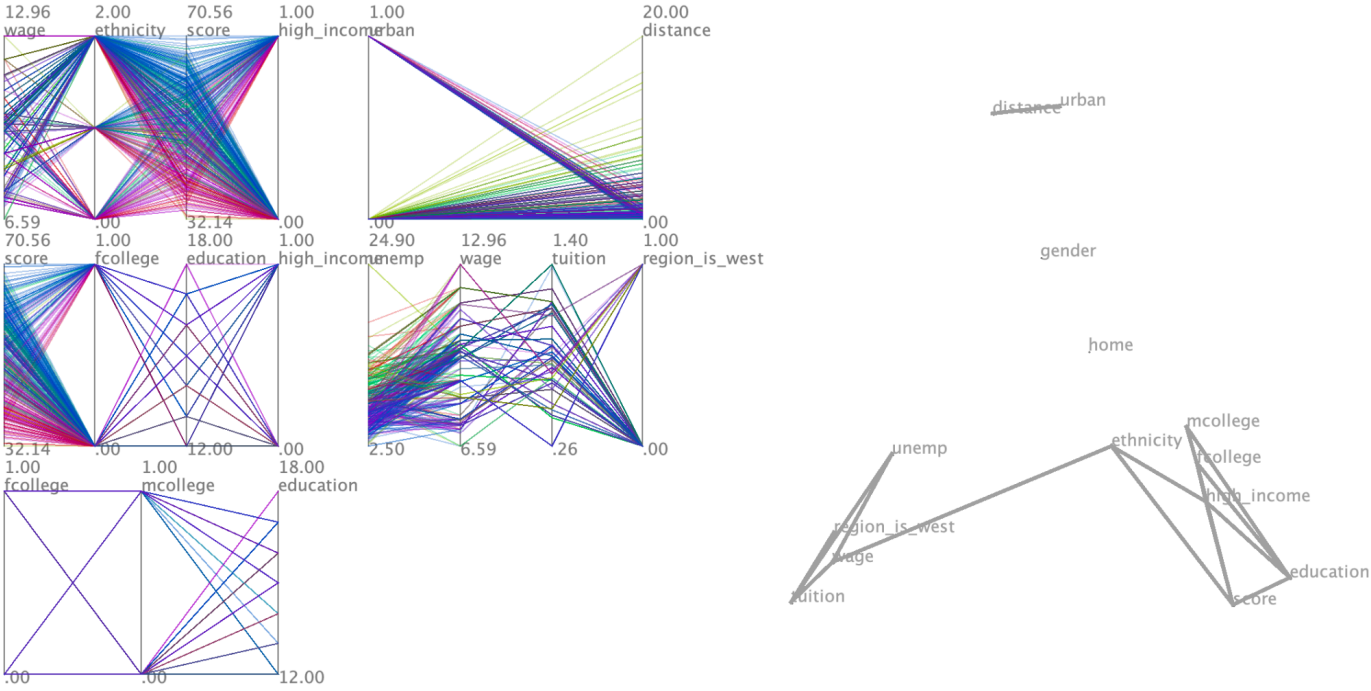


図9：解析結果(クラスタ数5)

＜考察＞
図9を見ると、出力されるPCPの数が5つとなり、かなりの要素が線分で結ばれている状態になった。この中で、まだ分析していない軸を取り上げてここでは考察を行う。
まず、**score**と**high_income**について、クラスタ数を**4**にしたのが図10である。
高所得である場合、低所得である場合に比べて低スコアの割合が低いことが赤い線分からわかる。また、高スコアの割合も高所得の方が高いことも紫の線から読み取れる。
つまり、親の所得と高校時代の成績は正の相関があることがこの結果から考えられる。

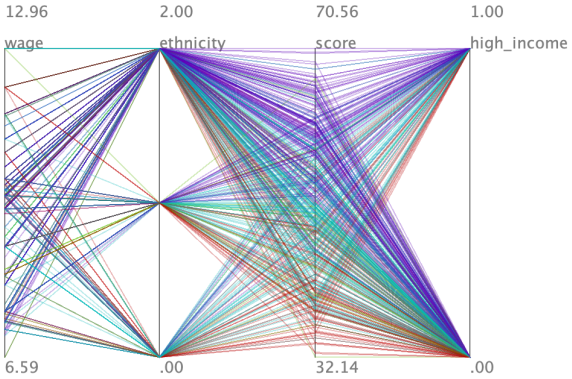


図10：図9左上のPCP(クラスタ数4)

次に、左側中段のPCPの図に注目すると、**score**と**fcollege**の関係性を分析できる。ピンクの線(=成績下位層)の割合が、父親が大学を卒業していない場合に、卒業している場合よりも高くなっていることがわかる。青い線に注目すると、成績上位層については、父親の学歴はそれほど影響していないことが考えられる。つまり、父親が大学を卒業していないことは、子供の成績に影響する可能性があることが考えられる。

これ以上右スライダーを動かすと線分が全て繋がるので、次からはcullingを調整して分析する。

⑥**culling**スライダを動かして変化を見る
Cullingのスライダを動かした上で右スライダーを動かして、新しい軸が出てくるまで調節する。
結果は図11のようになった。

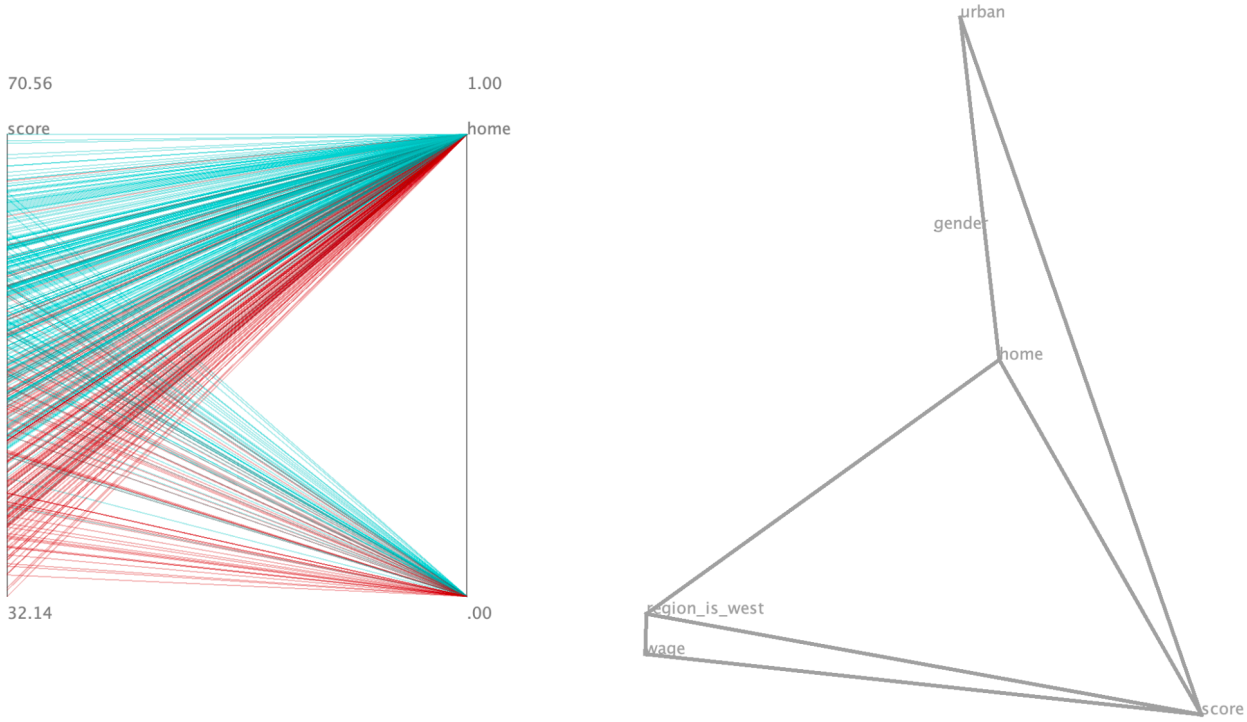


図11：解析結果(クラスタ数2)

＜考察＞
今回は**score**と**home**の軸でPCPが出力された。
図11のクラスタリングにより、家を所有している家庭の子供の方が、家を所有していない家庭の子供よりも成績の良い割合が多い(=水色の線の割合が多い)ことが読み取れた。
また、クラスタ数を**4**に変えた図12より、成績最上位層(=紫色の線) の割合については、家を所有している家庭の子供の方が圧倒的に割合が高いことが読み取れた。
これらのことから、家を所有しているかどうかは子供の成績に大きく影響することが考えらえる。
これは、経済的に豊かで余裕のある家庭は家を所有しており、子供の教育にもより多くのお金をかけることができているためだと考えた。

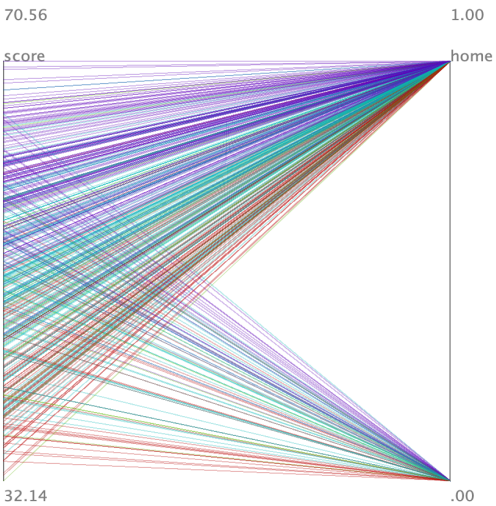


図12：PCPの様子(クラスタ数4)

⑦**culling**を最大まで動かして変化を見る
Cullingを最大まで動かしたところ、**gender**と**score**の2つのパラメータのみ残った。
結果は図13のようになった。

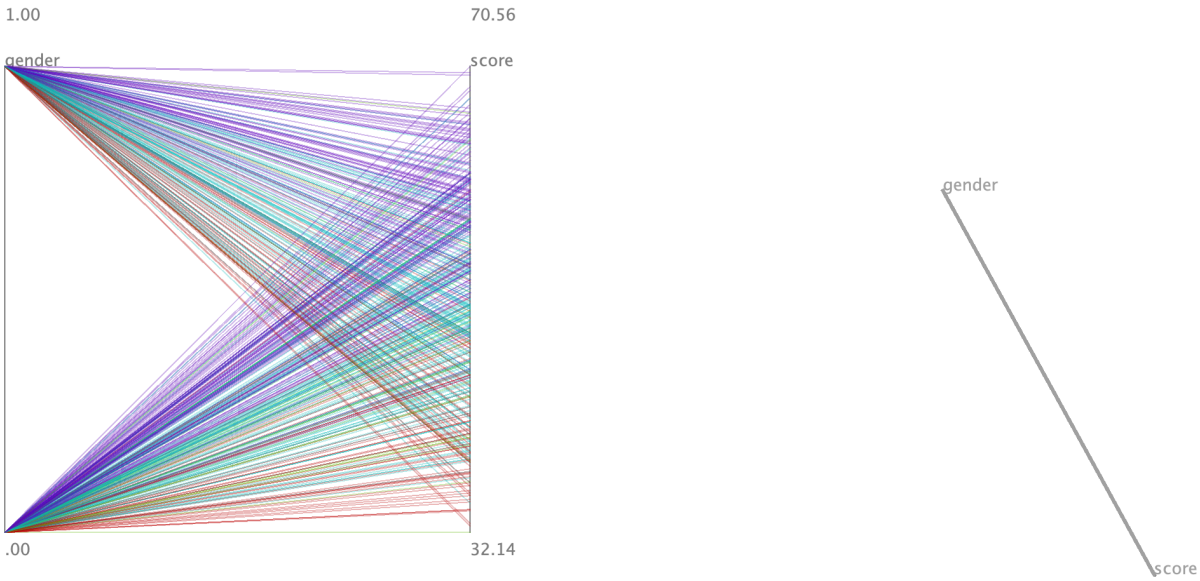


図13：解析結果(クラスタ数4)

＜考察＞
Cullingを最大まで動かした時に最後に残った軸は**gender**と**score**だった。cullingを動かして最後まで残ったということは、この**2**つの相関は非常に弱いと予測できる。
これを踏まえた上で、クラスタ数を**4**に設定すると、図13の左側のPCPが得られた。
Genderが1(男性)、0(女性)であっても、クラスタリングした線分の色に大きな差はない。全ての色の割合が、男性の場合と女性の場合でほぼ等しくなっていることが視覚的に確認できる。
つまり、性別による成績の差は見られないことが推測できる。

①～⑦がスライダーを動かした際に得られる結果の分析の全てである。
しかし⑤で述べたように、**fcollege**, **mcollege**, **score**, **education**, **high_income** がすごく近くにあるのに軸として現れてきていないのでこれを軸にした結果を表示したいと考えた。しかし、2値同士のデータを軸にすると線分の数が激減し、得られる情報が極端に少ない。そのため、**2**値の**high_income**を消去し、**fcollege**, **mcollege**, **score**, **education**の4つにデータを絞ってみて新たに分析をかけてみることにした。データを絞った後のファイルは図14のようになった。
これを用いて以降の分析を行う。

| 1 | Numeric | Numeric | Numeric | Numeric |
|---|---|---|---|---|
| 2 | score | fcollege | mcollege | education |
| 3 | 39.1500015 | 1 | 0 | 12 |
| 4 | 42.2200012 | 0 | 0 | 12 |
| 5 | 50.9300003 | 0 | 0 | 13 |
| 6 | 68.5800018 | 1 | 1 | 16 |
| 7 | 70.0999985 | 1 | 0 | 17 |
| 8 | 49.6500015 | 0 | 0 | 15 |
| 9 | 57.7400017 | 0 | 0 | 16 |
| 10 | 60.9300003 | 0 | 0 | 14 |
| 11 | 65.3099976 | 1 | 0 | 16 |
| 12 | 45.1399994 | 0 | 0 | 12 |
| 13 | 40.2799988 | 0 | 0 | 12 |
| 14 | 50.8100014 | 0 | 1 | 16 |
| 15 | 50.9500008 | 0 | 0 | 12 |
| 16 | 54.2599983 | 0 | 0 | 14 |

図14：データ削除後のcsvファイル

⑧データを絞った上での初期状態
データを図14のように絞った状態で、再び分析を行った。
図15は、スライダーを動かしていない初期状態の様子である。

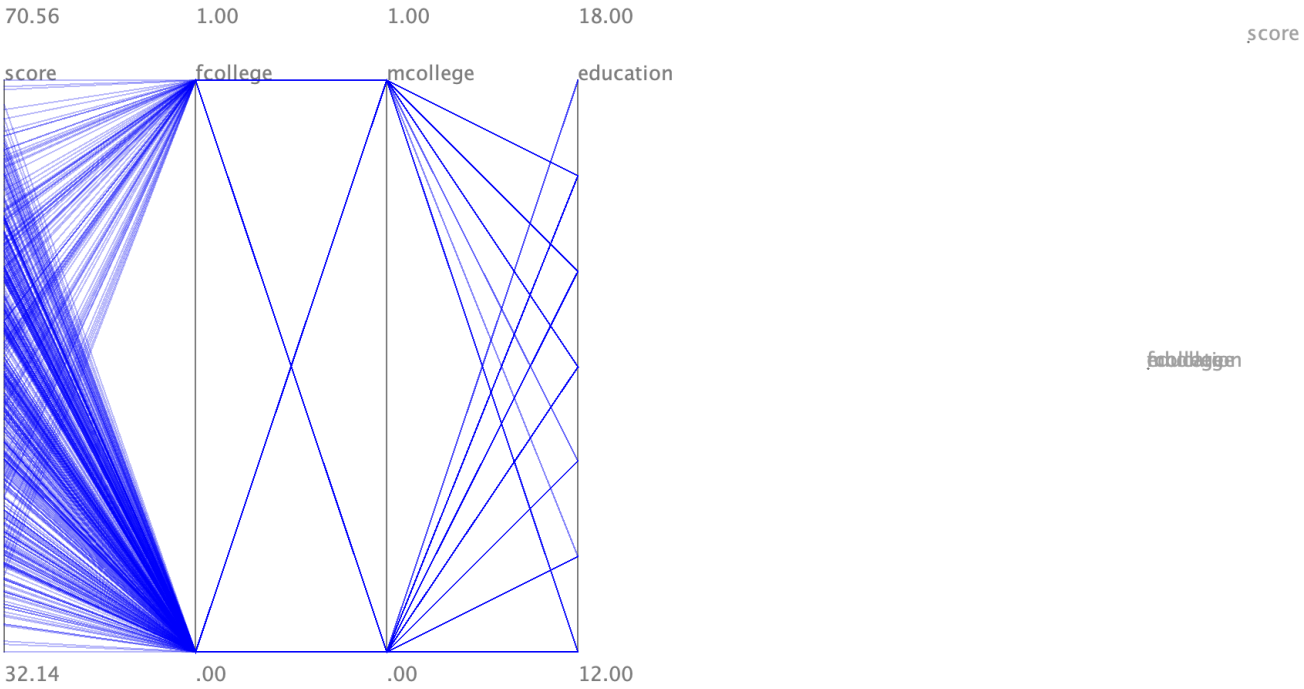


図15：解析結果(クラスタ数1)

＜考察＞
Score以外の要素(**fcollege**, **mcollege**, **education**)が重なるように散布図において分布していた。これらの要素は非常に相関が強いことが散布図から読み取れる。つまり、両親が大学に行っているかどうかは子供の教育年数に非常に大きな影響があることがわかる。

初期状態では、⑤において分析した**score**と**fcollege**の軸のPCPが出力された。これも、クラスタ数を**3**にしてみると、より分かりやすくなった。
父親が大学を卒業していないことは、子供の成績に影響することがこのグラフから読み取ることができた。

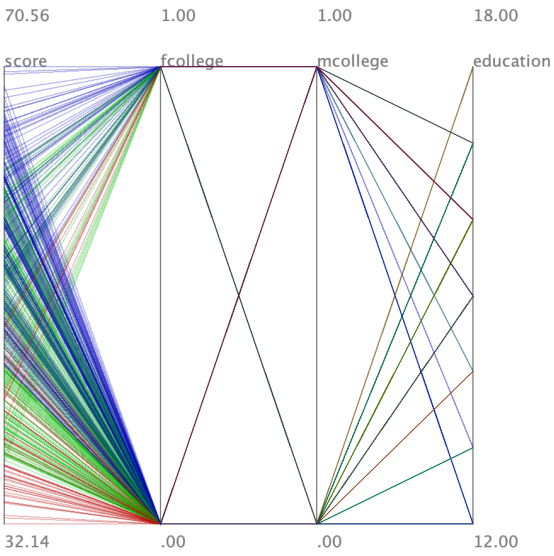


図16：PCPの様子(クラスタ数3)

次に、右側のスライダやcullingのスライダを調整して出力されるPCPや散布図を観察したが、**score**と**mcollege**を軸としたPCPを得ることができなかった。
しかし、⑧の散布図では**score**と**mcollege**は散布図上でほぼ重なっており、これも**score**と**mcollege**の関係と同様に相関関係があるのではないかと考えた。
そのため、**fcollege**のデータを消去した上で、再度分析を行い、出力について分析することにした。

### ⑨**fcollege**のデータを削除した上で分析する

Score, mcollege, education の3つのデータで分析を行った結果、図17のようになった。

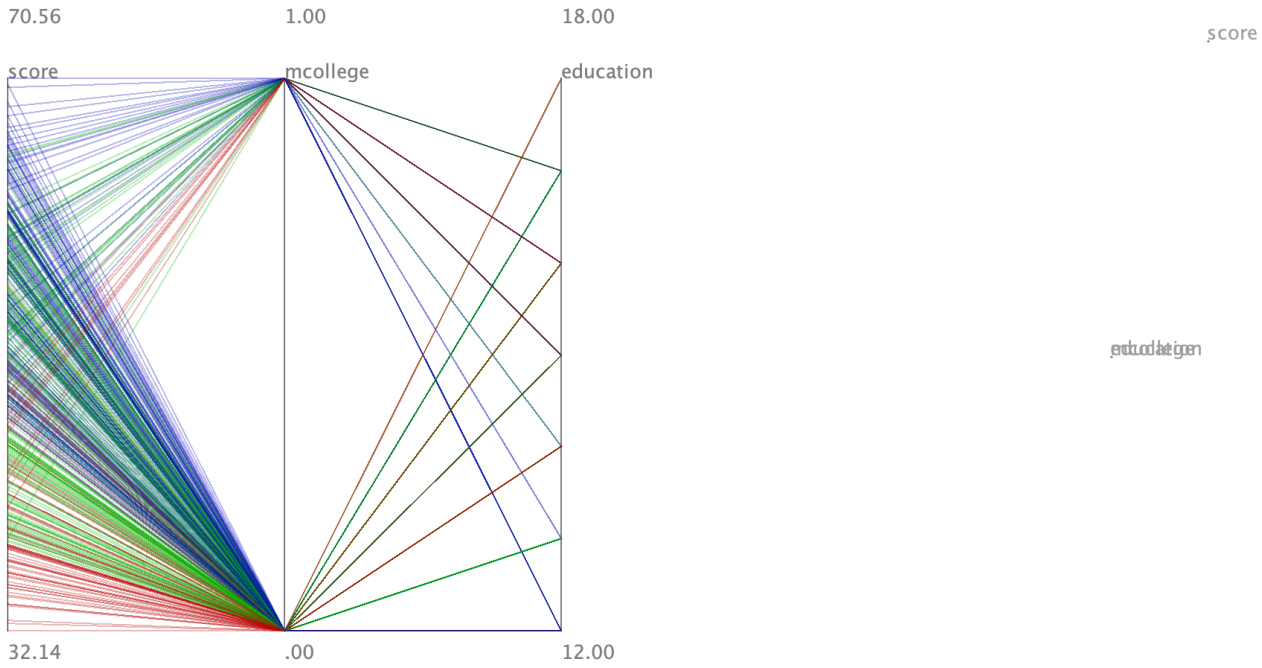


図17：解析結果(クラスタ数4)

＜考察＞
**fcollege**を削除する前と同様に、**score**以外の要素(**mcollege**, **education**)は強い相関があることが図17の散布図から読み取れる。
また、図17のPCPより、母親が大学を卒業している場合、卒業していない場合よりも高い成績を取る割合が高いことが青い線からわかる。反対に、母親が大学を卒業していない場合、卒業している場合よりも低い成績を取る割合が高いことが赤い線から読み取れる。
つまり、母親の学歴は、子供の成績に非常に大きく関係していることがこのグラフから読み取れる。
父親の学歴の場合とはわずかに異なる結果となったが、まとめると、両親の学歴は子供の成績に非常に大きく関係していると言える。

4. 感想

クラスタ数を変更してみると、ぱっと見ではわからなかった傾向や、より細かな情報が得られたのでとても驚いた。
今回は2値やカウンタの項目が半分以上あり、せっかくデータが約500個あったのにPCPの図で線がほぼ重なってしまって情報が減ってしまったのが残念だった。連続で質的な変数をたくさん入れると、より情報量の多い分析ができるのではないかと思った。
説明書に書いてあった通り、散布図で近くに分布している要素同士のPCPが現れないことがあるのがもどかしいなと思った。今回はデータをその部分だけ取り出して分析を行ったが、軸を自由に選べるようにもなるといいなと感じた。

5. 参考文献

R: College Distance
https://vincentarelbundock.github.io/Rdatasets/doc/AER/CollegeDistance.html