

Data Mining: US Census income data

Pol Casacuberta
He Chen
Eric Hurtado
Tommaso Patrì
Alexandru-Ilie Popa



Index

1. Overview
2. Data Mining process
3. Descriptive analysis
4. Preprocessing
5. PCA
6. Clustering
7. Profiling
8. Conclusions

Overview

Goal

- Identify patterns, relationships, and correlations within the data and draw conclusions about the factors that may impact income.

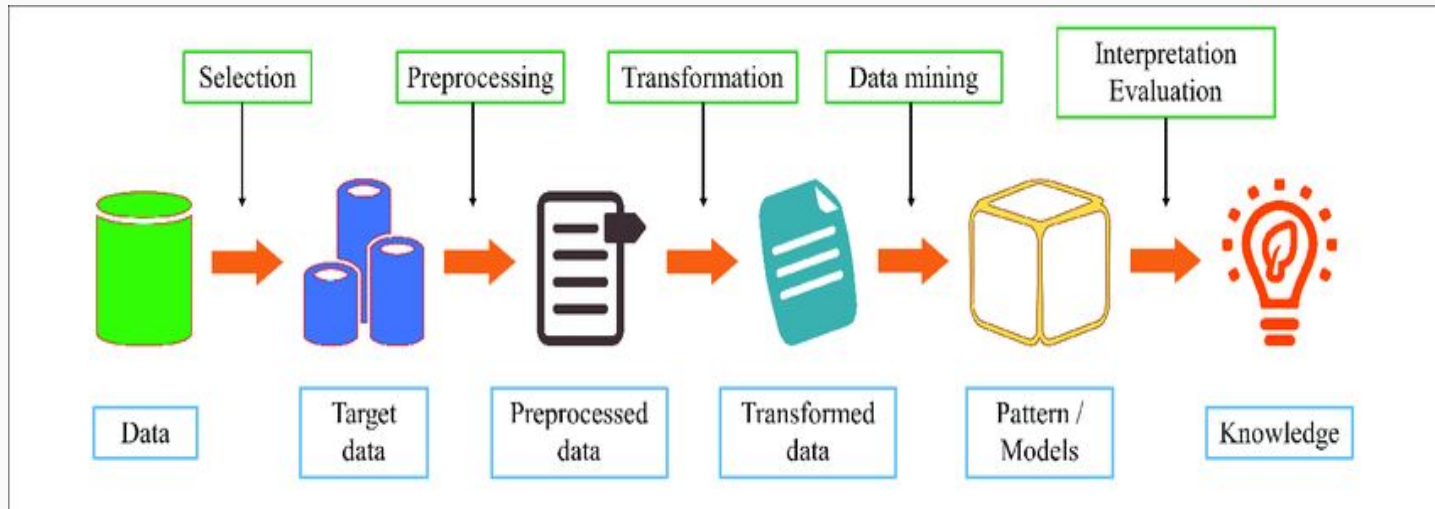
Original Dimension

- 199,523 individuals (rows)
- 42 features (columns)

After Sampling and Data Selection

- 20,000 individuals (rows)
- 28 features (columns)

Data Mining Process

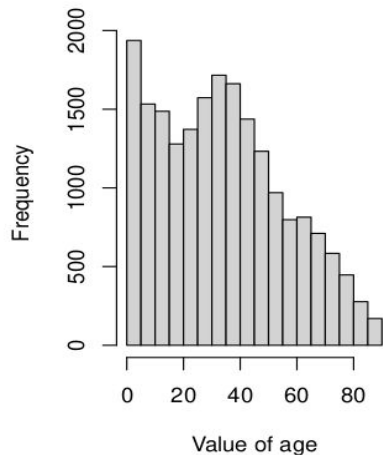


Descriptive Analysis

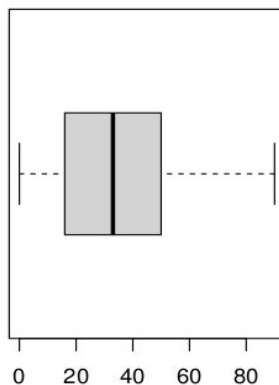
Univariate Analysis

Variable 1 : age

Histogram of age



Boxplot of age

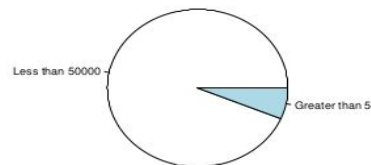


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	16	33	34.41035	50	90	22.11444	0.6426684	0

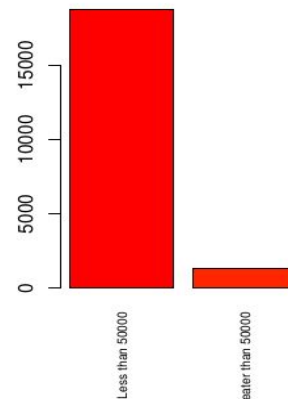
Table 6.3. age extended Summary Statistics.

Variable 28 : income

Pie of income



Barplot of income

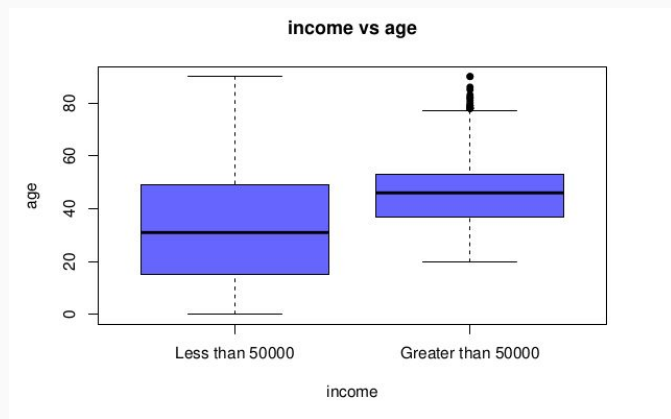
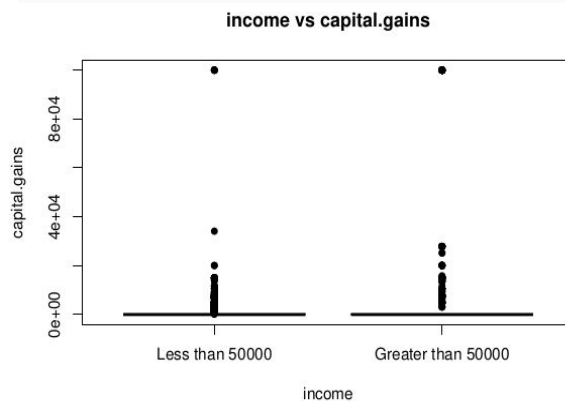
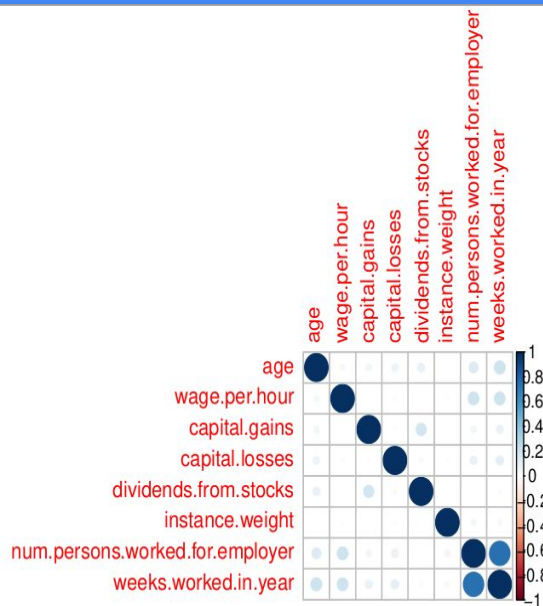


Number of modalities: 2

income	Frequency	Proportion
Less than 50000	18728	0.9364
Greater than 50000	1272	0.0636

Table 6.40. income frequency and proportion table.

Bivariate analysis



Preprocessing

1. Feature selection

- Remove features that do not contribute any information to the topic

2. Prepare and cleaning data

- Convert missing data value
- Categorical value to factors
- Set levels

Preprocessing

3. Missing data: all categorical data

Group	Description	Strategy
1	less than 10%	try to impute or create new category
2	Above 50%	remove

Preprocessing

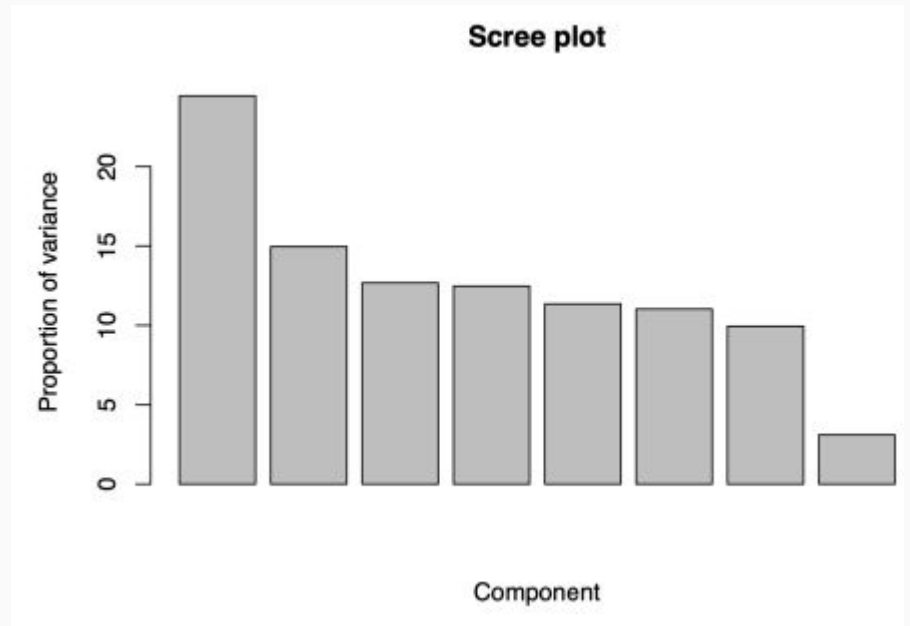
Group 1: hispanic origin, country of birth self

Group 2: class of worker, major industry code, major occupation code, live in this house 1 year ago ...

Group 3: enrolled in edu inst last wk, member of a labor union, ...

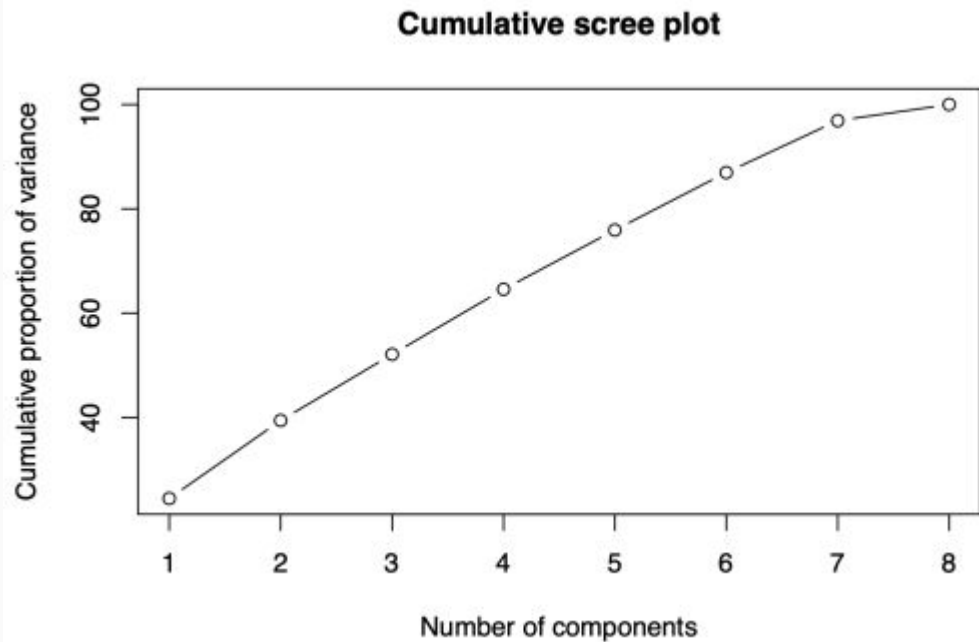
PCA

- Similar amount of variance



PCA

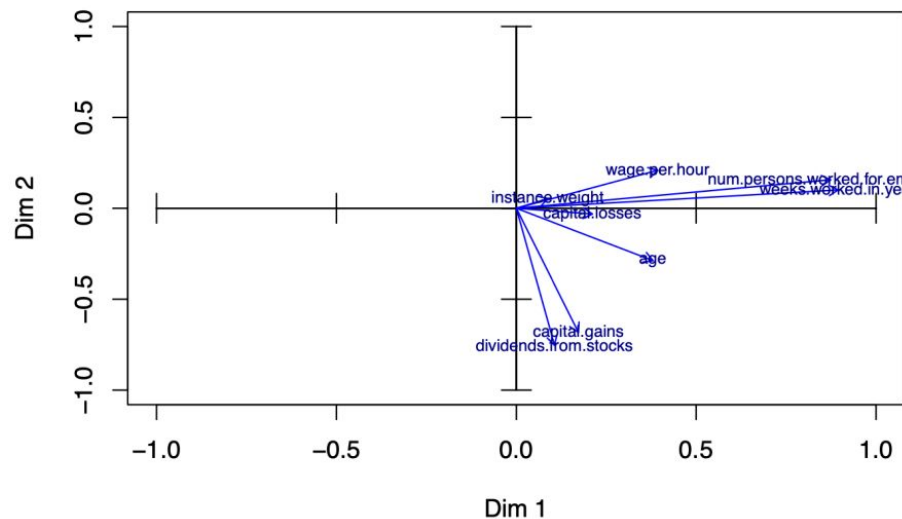
- 5 components $\approx 80\%$
- We choose 5 PCs to represent our data



PCA

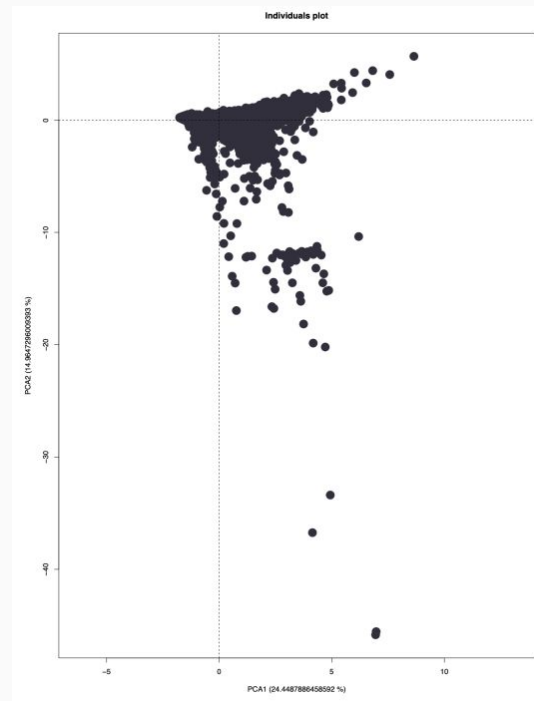
- Most important in y-axis:
 - Dividends from stocks
 - Capital gains
- Most important in x-axis:
 - weeks worked in a year
 - number of persons worked for the employer
- Positive correlation between:
 - capital gains, dividends from stocks and age
- Negative correlation y-axis:
 - Wage per hour and age

Projection of numeric variables in X: 1, Y: 2



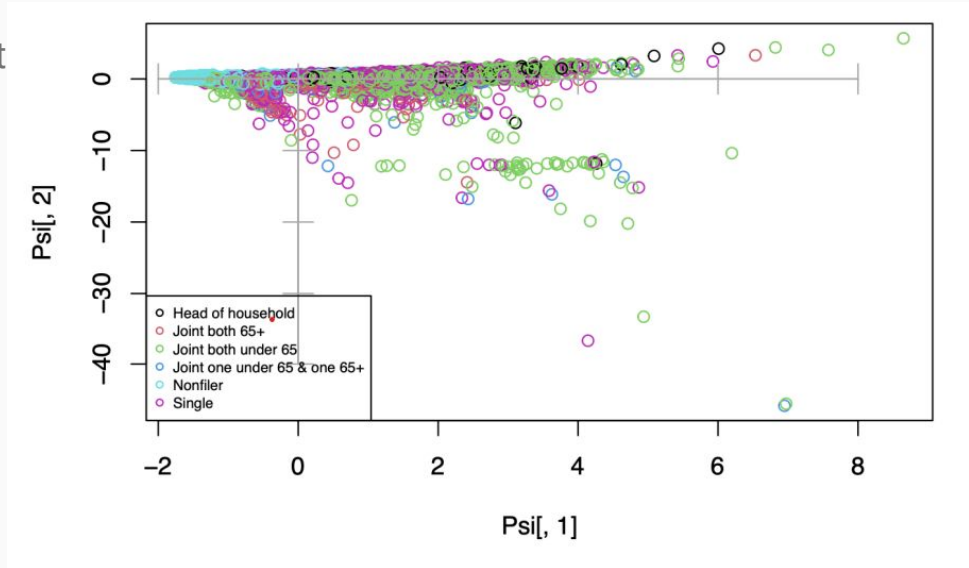
PCA

- Most observations are clustered around the center of the plot
- Some outliers



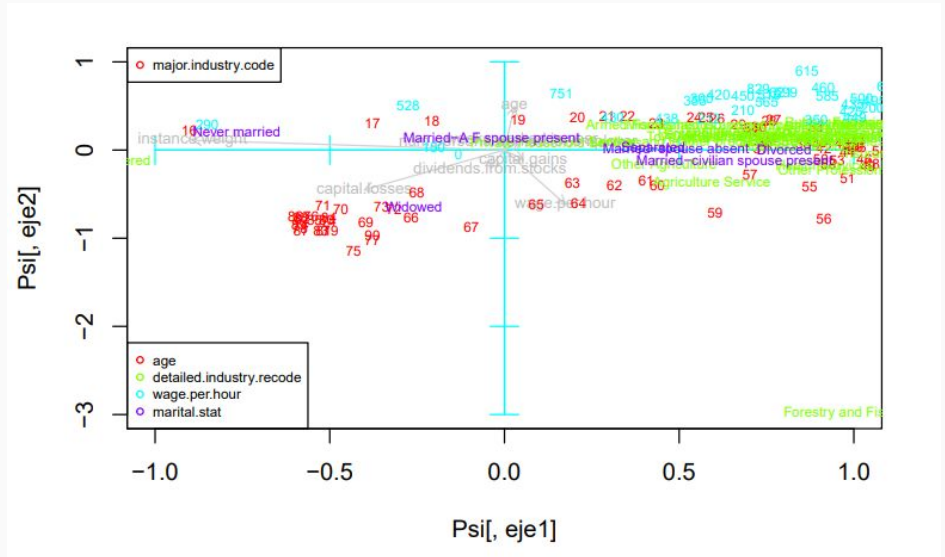
PCA

- detailed.household.and.family.stat
- Nonfiler: identifies taxpayers who have not filed a federal or state individual income tax return for the tax year under review



PCA

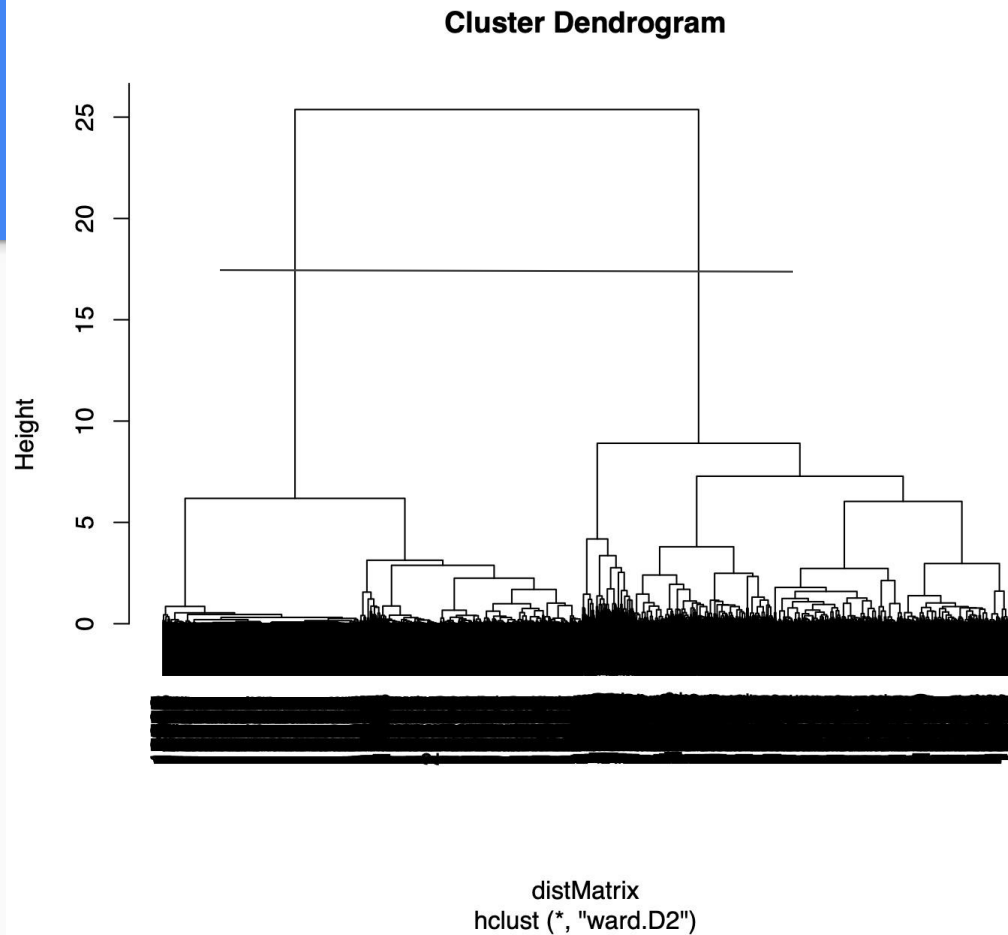
- Centroids for the factors with respect to the principal components
- Widowed: their ages tend to be older
- Capital losses seem to be highly related with higher ages



Clustering

method="ward.D2"
metric = "gower"

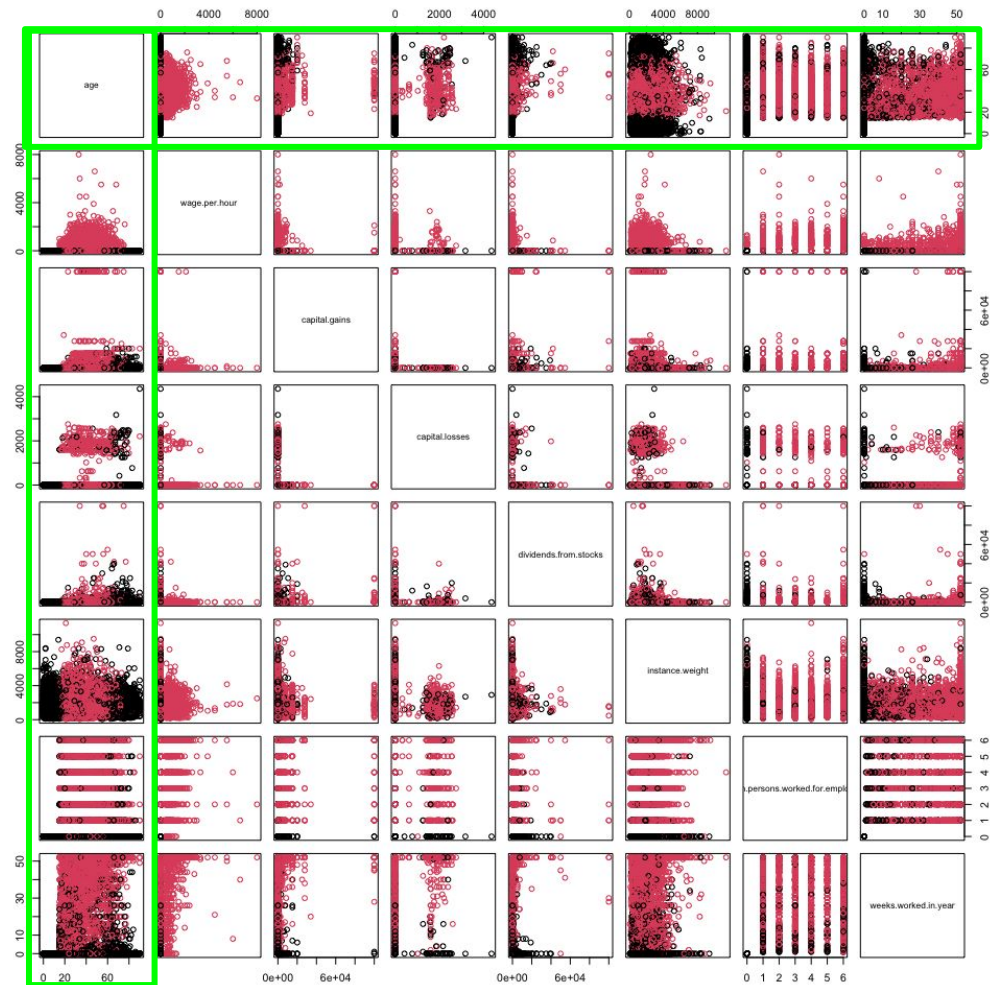
1	2
9923	10077



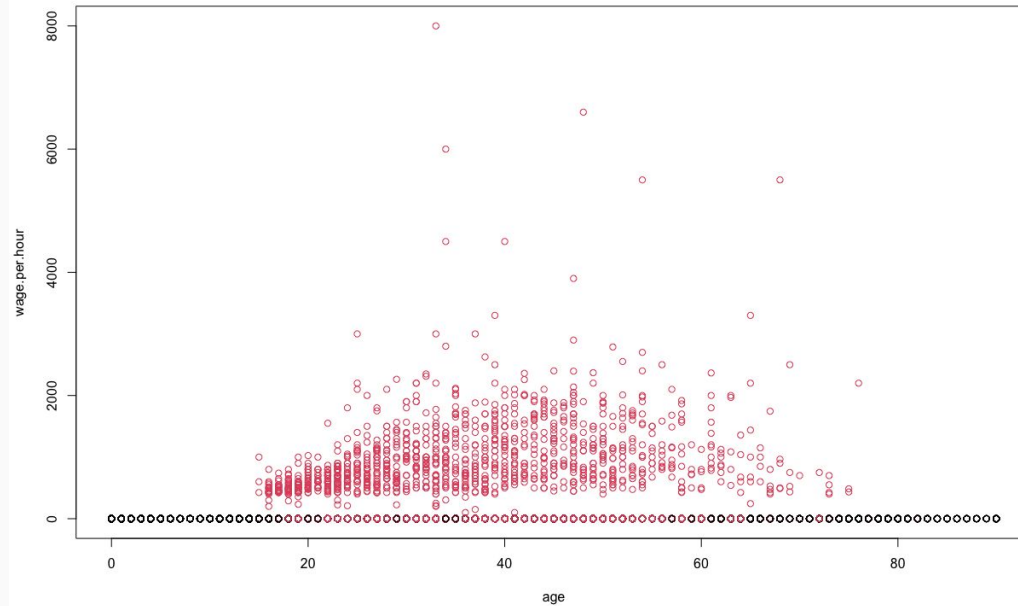
Clustering

Cluster 1 = Age: 0-17 -> 65-90

Cluster 2 = Age: 18-64

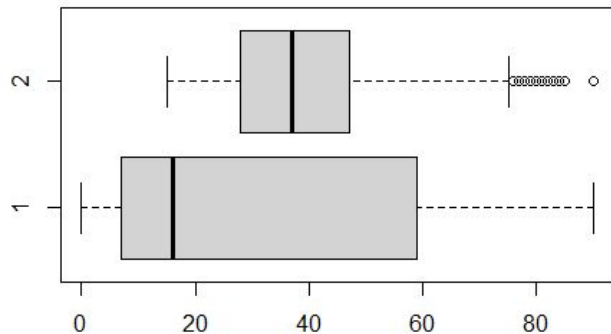


Clustering

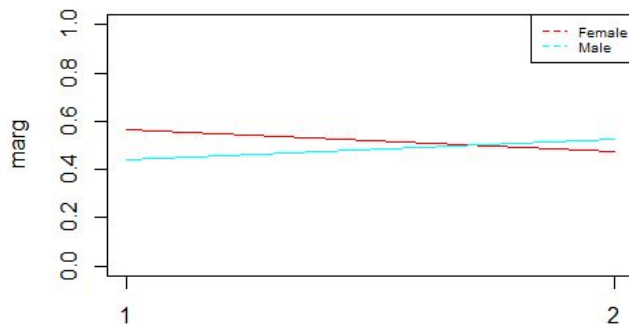


Profiling

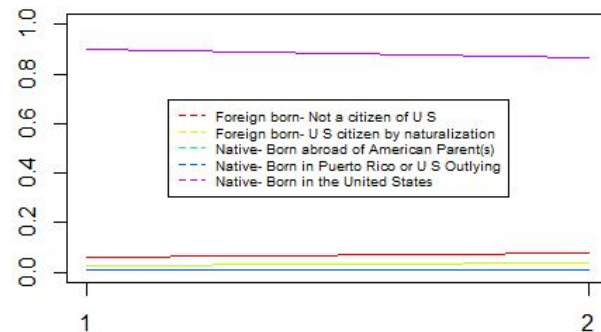
Boxplot of age vs Class



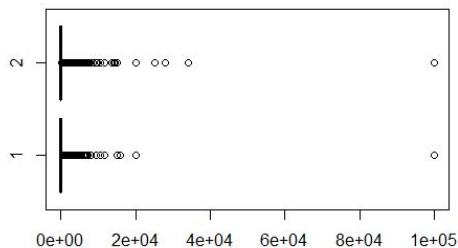
Prop. of pos & neg by sex



Prop. of pos & neg by citizenship

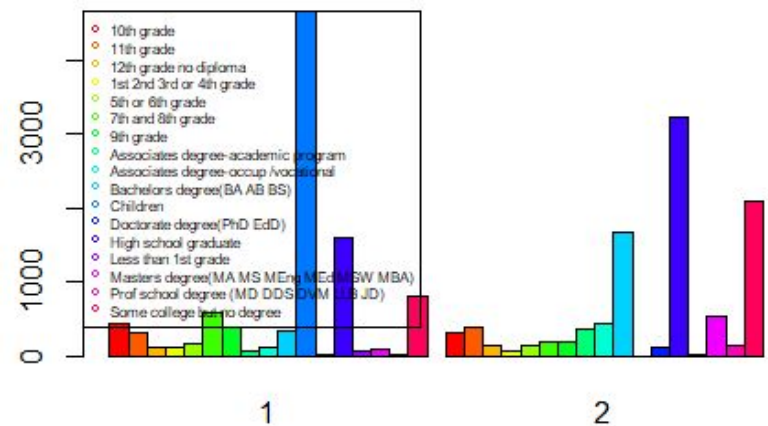


Boxplot of capital.gains vs Class



Socially, the American society do not have great difference

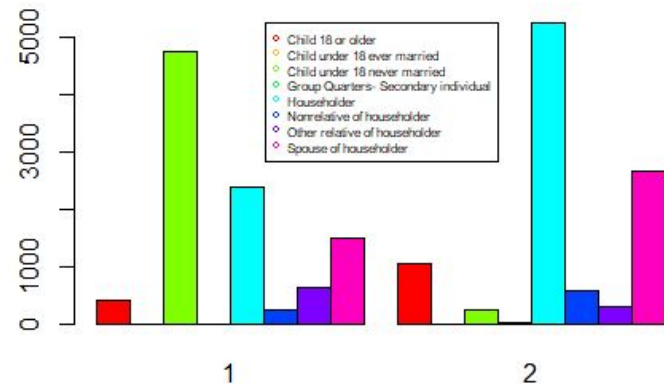
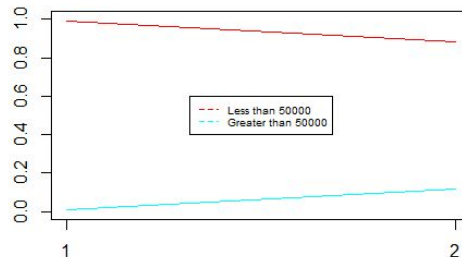
Profiling



Education

There is a generational difference

Prop. of pos & neg by income



Household

Final conclusions

- Two distinct clusters: Workers and non-workers
- Dividends from stocks and capital gains
- Some Categorical variables such as industry are predominantly “not considered”
- Income, individuals with greater than \$50,000 present less variability with respect to age
- Individuals who identify as white have a higher wage per hour compared to other races
- The grand majority of individuals have an income of less than \$50,000 a significant amount of them are either in the armed forces or are children