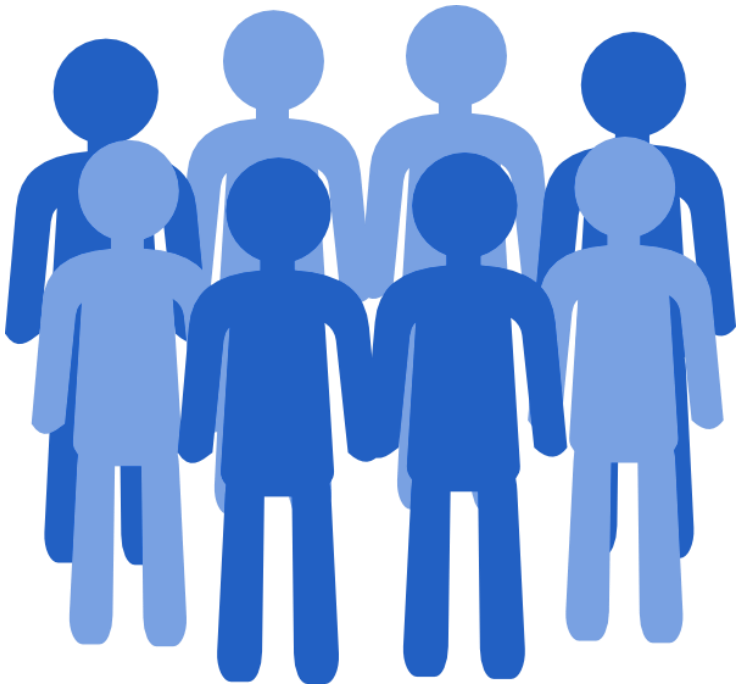
American Community Survey

Marta Cortés, Elianne Mora & Rafaela Becerra



1. Introduction

The main goal of the present work is to identify groups of counties with common characteristics based on the information selected, which has been catalogued as relevant and fundamental. We will use some descriptive analysis to determine the shape of data distributions, the presence of hidden groups and outliers, and performed clusters of counties using unsupervised tools.

This dataset is part of the U.S. Census Bureau's *American Community Survey*, which is a nationwide, continuous survey that collects data each year in order to get useful information about the communities.

Every year, over 3.5 million households are contacted across the US to participate in the ACS, by mail or courier, this means that about 1 out of 38 households are part of the survey (Figure 1).

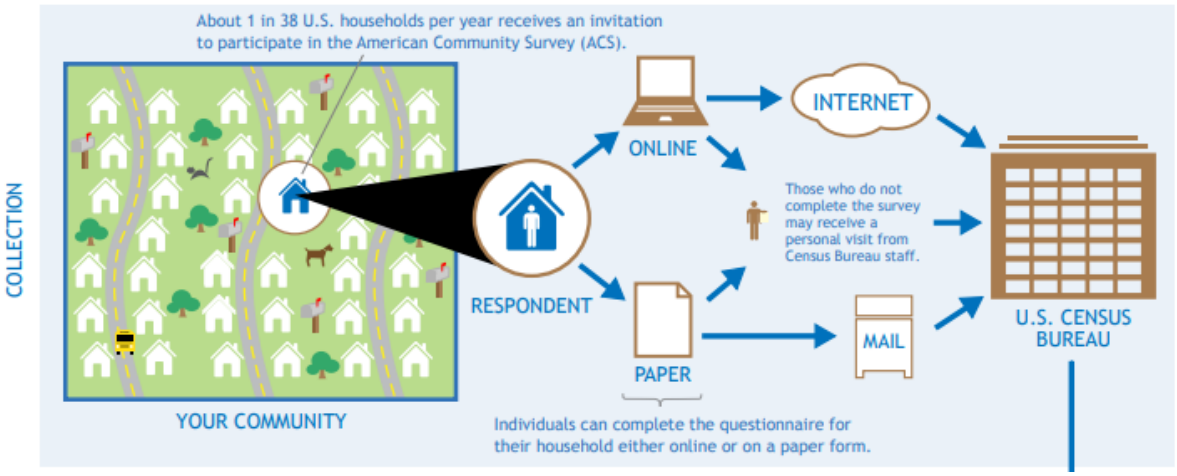


Figure 1. Collection diagram taken from ACS Information Guide (2017)

Since the data is collected annually, the Census Bureau presents in various ways this data, in a yearly basis for areas with populations of 65,000+, and as a compound of 60 months for data of all the areas.

The importance of this survey lies in the periodicity with which it is carried out because it involves annual surveys that present important information that would instead be collected in censuses with a periodicity of 10 years. Therefore, it provides imperative information to make decisions at all levels, on time (Figure 2).

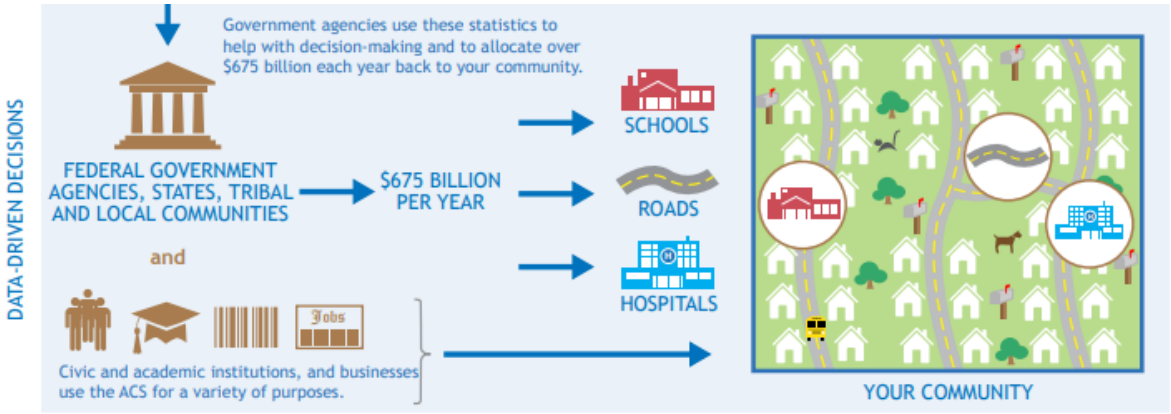


Figure 2. Data driven decisions of the ACS, taken from ACS Information Guide (2017)

The present dataset is part of the data products elaborated from the ACS of the 5 years estimates data for each county (3220) in the US for 2017 and which contains compare derived measures such as percentages, means, medians, and rates of population totals. The total variable set of 36 variables from the two products selected DP03 and DP05, are described in the following Table:

Name of the variable	Description	Type
State	State, DC, or Puerto Rico	Categorical
County	County	Categorical
TotalPop	Total population	Continuous
Men	Number of men	Continuous
Women	Number of women	Continuous
Hispanic	Percentage of population that is Hispanic/Latino	Continuous
White	Percentage of population that is white	Continuous
Black	Percentage of population that is black	Continuous
Native	Percentage of population that is Native American or Native Alaskan	Continuous
Asian	Percentage of population that is Asian	Continuous

Pacific	Percentage of population that is Native Hawaiian or Pacific Islander	Continuous
VotingCitizen	Number of citizens that vote in county	Continuous
Income	Median household income ($)	Continuous
IncomeErr	Median household income error ($)	Continuous
IncomePerCap	Income per capita ($)	Continuous
IncomePerCapErr	Income per capita error ($)	Continuous
Poverty	Percentage under poverty level (*)	Continuous
ChildPoverty	Percentage of children under the poverty level	Continuous
Professional	Percentage employed in management, business, science, and arts	Continuous
Service	Percentage employed in service jobs	Continuous
Office	Percentage employed in sales and office jobs	Continuous
Construction	Percentage employed in natural resources, construction, and maintenance	Continuous
Production	Percentage employed in production, transportation, and material movement	Continuous
Drive	Percentage commuting alone in a car, van, or truck	Continuous
Carpool	Percentage carpooling in a car, van, or truck	Continuous
Transit	Percentage commuting on public transportation	Continuous
Walk	Percentage walking to work	Continuous
OtherTransp	Percentage commuting via other means	Continuous

WorkAtHome	Percentage working at home	Continuous
MeanCommute	Mean commute time (minutes)	Continuous
Employed	Number of employed (16+)	Continuous
PrivateWork	Percentage employed in private industry	Continuous
PublicWork	Percentage employed in public jobs	Continuous
SelfEmployed	Percentage self-employed	Continuous
FamilyWork	Percentage in unpaid family work ('housewife' tasks)	Continuous
Unemployment	Unemployment rate (Percentage)	Continuous

Table 1. Description of variables.
(*) A family is considered to be officially in poverty if their pre-tax income is below a threshold set by the current value of three times a minimum food diet in 1963, adjusted by family composition (Figure 3).

Size of family unit	Weighted average thresholds	Related children under 18 years								
		None	One	Two	Three	Four	Five	Six	Seven	Eight or more
One person (unrelated individual):	12.488									
Under age 65	12.752	12.752								
Aged 65 and older	11.756	11.756								
Two people:	15.877									
Householder under age 65	16.493	16.414	16.895							
Householder aged 65 and older	14.828	14.816	16.831							
Three people	19.515	19.173	19.730	19.749						
Four people	25.094	25.283	25.696	24.858	24.944					
Five people	29.714	30.490	30.933	29.986	29.253	28.805				
Six people	33.618	35.069	35.208	34.482	33.787	32.753	32.140			
Seven people	38.173	40.351	40.603	39.734	39.129	38.001	36.685	35.242		
Eight people	42.684	45.129	45.528	44.708	43.990	42.971	41.678	40.332	39.990	
Nine people or more	50.681	54.287	54.550	53.825	53.216	52.216	50.840	49.595	49.287	47.389

Source: U.S. Census Bureau.

Figure 3. Poverty Thresholds for 2017 by Size of Family and Number of Related Children Under 18 Years.

2. Data Processing and Descriptive Analysis

2.1. Selection of the Variables

The DP03 and DP05 products contain categorical variables that divide the population. Since these are variables that are complementary, a selection of variables has been performed based on its importance.

In order to identify the composition for the total population, a calculation for all the indicators have been done, based on the information provided. The results are presented in Table 2. Also, the variables have been categorized, in order to identify which ones could be omitted from the analysis, resulting in 21 variables selected that are marked in orange.

In the case of gender, the variable women has been selected since it is considered that the role of women in the United States has changed dramatically over the past few decades, not only there are more women than men, also, they are stepping up to lead the country. They are majority at the public office, and a record-high percentage of women are serving in Congress; additionally, they are Nearly Half the Labor Force (Catalyst, 2019). However, there are substantial inequalities, for example, according to the Center of American Progress, the payment of women is 0.77 cents from what men make. Consequently, because of these disparities, the number of women can influence the behavior of the counties and must be considered for the analysis.

Moreover, from the variables of employment indicators, we have selected the rate of unemployment since it is a broader measure of underutilization in the labor market and includes those looking for work that are part of the labor force, and those working part-time, but that would like to have a full time job. This is a variable with higher implications than the number of people that have full time jobs that are regulated, which is represented by the rate of employment.

From the category of race, we have selected just three main categories: white, black and Hispanic populations since more than 86% of the total population are within these groups. The same consideration has been taken into account for the type of work category, since the groups “PrivateWork”, “PublicWork”, and “Selfemployed” contain the 99.83% of the population.

Additionally, the measure of income per-capita is considered to be a more general measure of the overall resources a person may receive in a certain area, because it involves the calculation of the personal income of the residents of a given area divided by the resident population. In contrast, the income is given by the median household resources and will not show the cases that are more deviated. Consequently, as the variable income can be better captured by the income per-capita, it has been selected.

According to the American Public Transportation Association, in the US, 45% of its residents have no access to public transportation. For every $1 invested in public transportation, it generates $4 in economic returns in this country; these facts show that enhancing and strengthening the public transportation system would lead to economic development, hence it is important to compare the behavior of this variable. Additionally, we have picked “drive” and “walk” as other means of commute to better understand the relationships, significance and how Americans get to work.

Furthermore, for the category “Labor”, defined as the classification of work, we have selected the three major categories, which are: “Professional”, “Service” and “Office”; we have also included “Construction”, because it comprises employees of an important industry in the US that has

presented positive trends in the last years and denote a residential market that it is booming (U.S. Construction Industry - Statistics & Facts, 2019). Therefore, this rate of workers could denote the construction industry's growth in each area.

The variables that were presented as one "MeanCommute" and "VotingCitizen" will be considered for the analysis, this is not the case for "WorkAtHome", as it only contains 4% of the population, can be omitted.

Category	Variable	Rate
Classification	County	3220
	State	52
Gender	Men	49.21%
	Women	50.79%
Population	VotingCitizen	70.77%
Eployment	Employed	46.73%
	Unemployment	6.76%
Race	Hispanic	18.48%
	White	60.81%
	Black	12.16%
	Native	0.65%
	Asian	5.24%
	Pacific	0.15%
Poverty	Poverty	14.93%
	ChildPoverty	20.61%
Labor	Professional	36.84%
	Service	18.16%
	Office	23.59%
	Construction	9.06%
	Production	12.34%
Transport	Drive	76.72%
	Carpool	9.21%
	Transit	4.91%
	Walk	2.69%
	OtherTransp	1.81%
Work at home	WorkAtHome	4.65%
Type of work	PrivateWork	79.71%
	PublicWork	14.10%
	SelfEmployed	6.02%
	FamilyWork	0.16%
Income	Income	$48,994.97
	IncomePerCap	$25,657.03
Time of transportation	MeanCommute	23.47453

Table 2. US rates for variables in data set.

2.2. Final data set

The data matrix that will be used for the following sections is denoted as X and contains 21 columns given by the number of variables detailed and 3220 rows, which are the number of counties in the United States, which will be treated as observations. The next tables (3 and 4) show some descriptive measures from the final set of variables. There are no missing values in the variables chosen, so no imputations or suppressions of observations were performed, conserving the 3220 rows.

Territorial classification	Variable	Count
1	County	3220
2	State	52

Table 3. Identifier variables

Quantitative variables	Variable	Units	Mean	Standard Deviation	Percentil 0	Percentil 25	Percentil 50	Percentil 75	Percentil 100	Histogram
1	Women	Number of people	50	165,216	35	5,554	12,994	33,594	5,126,081	
2	Hispanic	%	11.3	19.3	0	2.1	4.1	10	100	
3	White	%	74.9	23.1	0	63.5	83.6	92.8	100	
4	Black	%	8.68	14.3	0	0.6	2	9.5	86.9	
5	VotingCitizen	Number of people	71,310	210,869	59	8,442	19,699	50,366	6,218,279	
6	IncomePerCap	US dollars $	25,657	6,668	5,943	21,568	25,139	28,997	69,529	
7	Poverty	%	16.8	8.31	2.4	11.5	15.4	19.8	65.2	
8	Professional	%	31.5	6.52	11.4	27.2	30.5	34.9	69	
9	Service	%	18.2	3.74	0	15.8	17.8	20.2	46.4	
10	Office	%	21.9	3.17	4.8	19.9	22.1	23.9	37.2	
11	Construction	%	12.6	4.14	0	9.8	12.1	14.8	36.4	
12	Drive	%	79.6	7.66	4.6	77.3	81	84.1	97.2	
13	Transit	%	0.939	3.07	0	0.1	0.3	0.8	61.8	
14	Walk	%	3.24	3.89	0	1.4	2.3	3.82	59.2	
15	MeanCommute	Minutes	23.5	5.69	5.1	19.6	23.2	27	45.1	
16	Unemployment	%	6.67	3.77	0	4.47	6.1	8	40.9	
17	PublicWork	%	17.1	6.39	4.4	12.7	15.9	19.9	64.8	
18	PrivateWork	%	74.9	7.65	31.1	71.2	76.1	80.2	88.8	
19	SelfEmployed	%	7.77	3.86	0	5.2	6.8	9.2	38	

Table 4. Quantitative variables

There are some major deviations from the race components, this means that the composition per county vary substantially in some cases and is evident because there are counties which are characterized by their population. Additionally, the rate of poverty varies in 8.31 points, this can be considered high and denotes some disbalances in the repartition of resources through the populations - this can be attributed to the different levels of State funding and how resources are allocated throughout the counties by government offices. Also, the mean rate of use of public transportation is low, and there is a relatively low standard deviation; therefore, it can be said that, overall, the variable has low usage rates in most of the counties, which we can attribute to the poorly planned and established transportation system in the US. As we can see the histograms denote some asymmetry in almost all the variables, this point will be treated in the next sections.

Furthermore, one more variable has been added to the classification group in order to improve visualization of the counties. There are 4 regions which have been determined in the US in order to separate states by geographical zone: Northeast, Midwest, South, and West (U.S. Census

Bureau). Puerto Rico is a Free Associated State, which does not belong to any of this regions and will be assigned as one separate region.

2.3. Distribution of the Variables

2.3.1. Density and QQ-Plots

In the next two figures it is shown the density graphs and the QQ-plots for each quantitative variable in order to identify the symmetry and the probability distribution that they follow. The QQ-plots are the graphical visualization of the relation between the sample data of each variable (axis x), sort it in ascending order, and the theoretical quantiles calculated from a normal distribution (axis y), the deviations of the points given by each observation from a straight line should be minimal in order to assume that they came from a normal distribution.

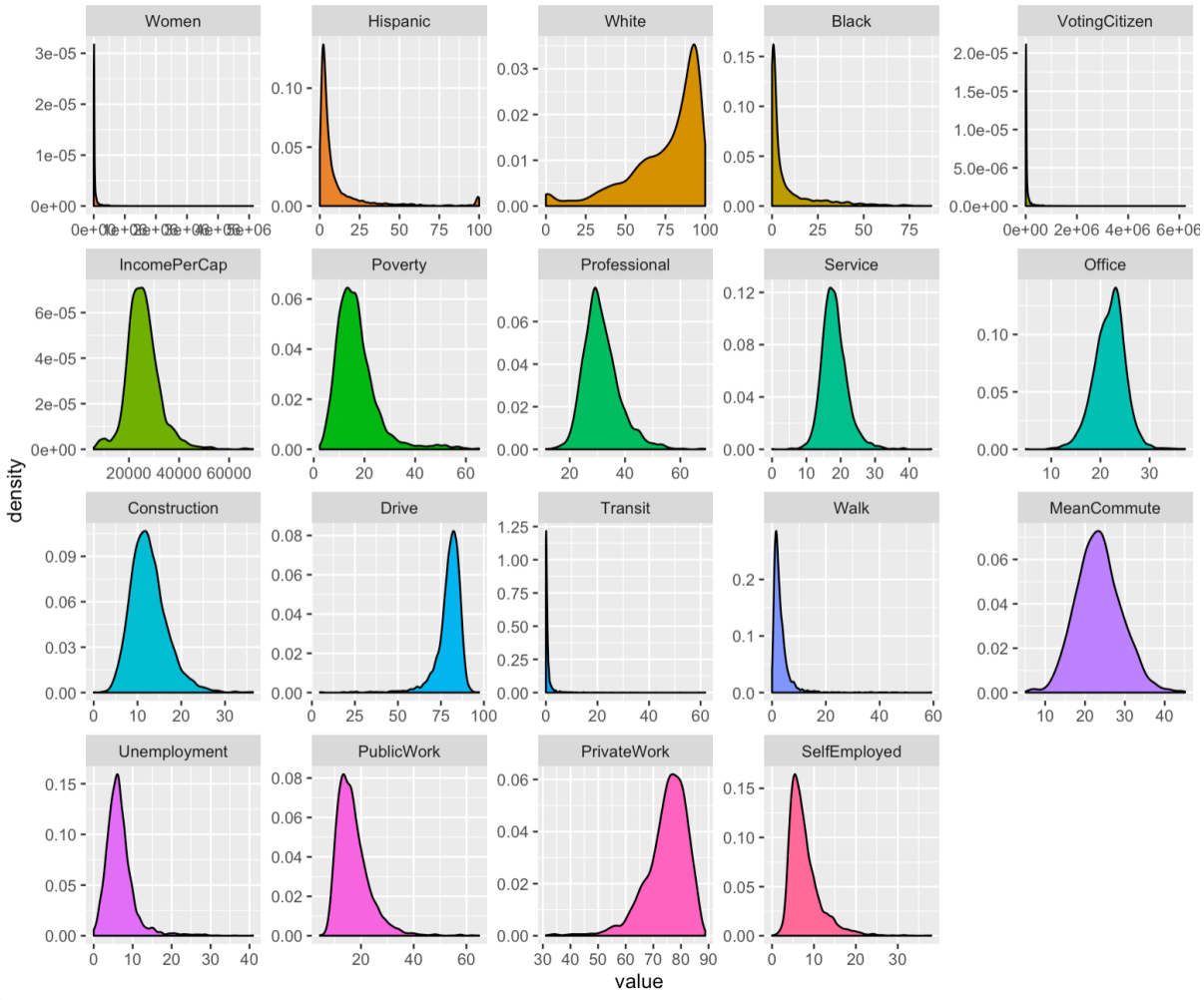


Figure 4. Density graphs of each variable

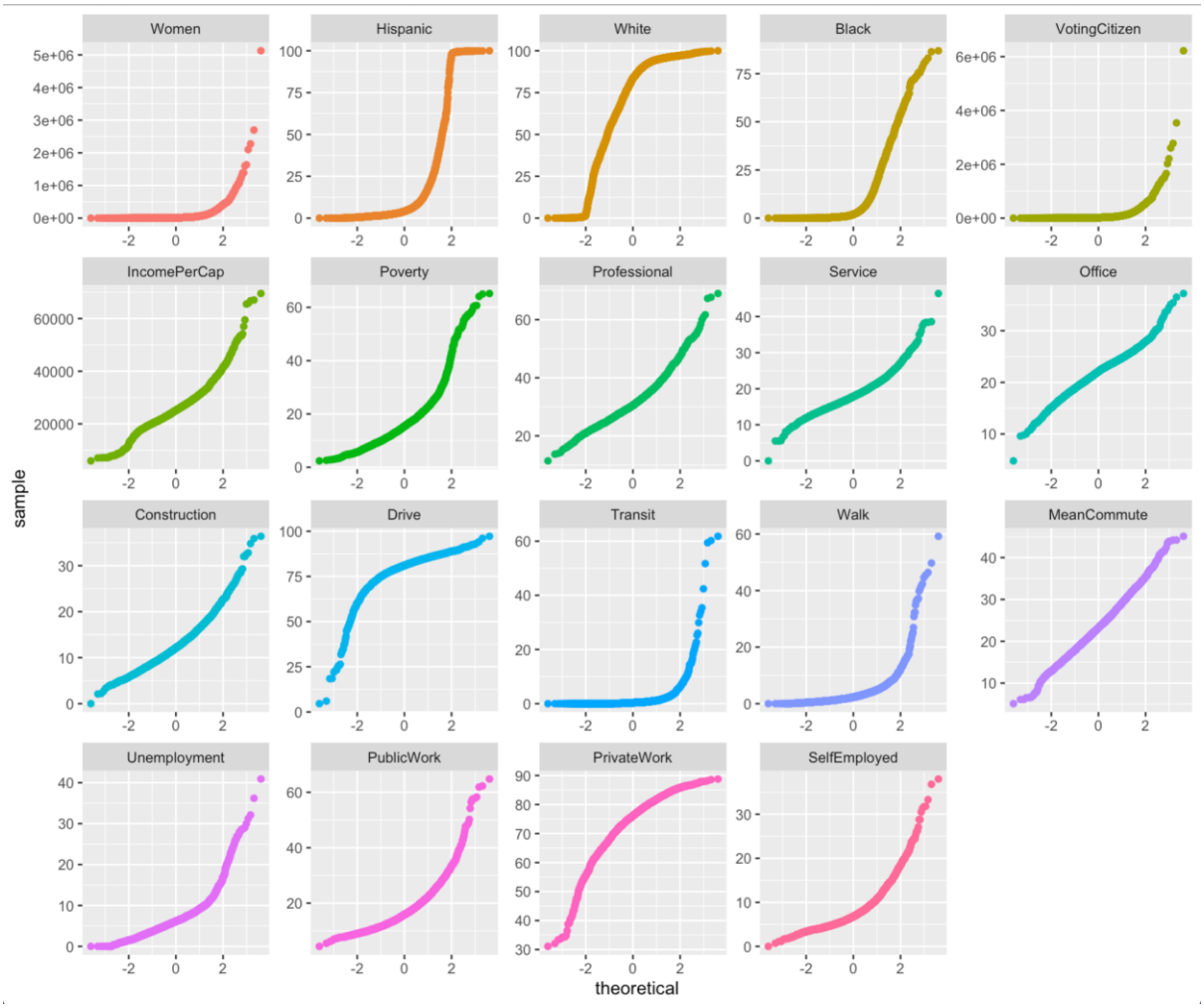


Figure 5. QQ-Plots of each variable

The most symmetric and linear variables are Income per Capita, Professional, Service, Office, Construction and Mean Commute. Although they can have the gaussian curve un-centered because a high variance or far outliers, we may say, they behave as a normal distribution, especially the Mean Commute, since this variable is already a mean of values.

Nevertheless, the variables Women, Hispanic, White, Black, Voting Citizen, Poverty, Drive, Transit, Walk, Public Work, Private Work, Self Employed have an exponential or gamma distribution as the curve is higher at some point close to the "y" axis and then it decreases quickly reaching zero or vice versa. Also, the data representation of the values doesn't fit with a linear relationship, but with a squaring relationship, creating a curve with positive tendency.

We can apply a transformation to these last variables in order to treat them as normally distributed.

2.3.2. BoxPlots

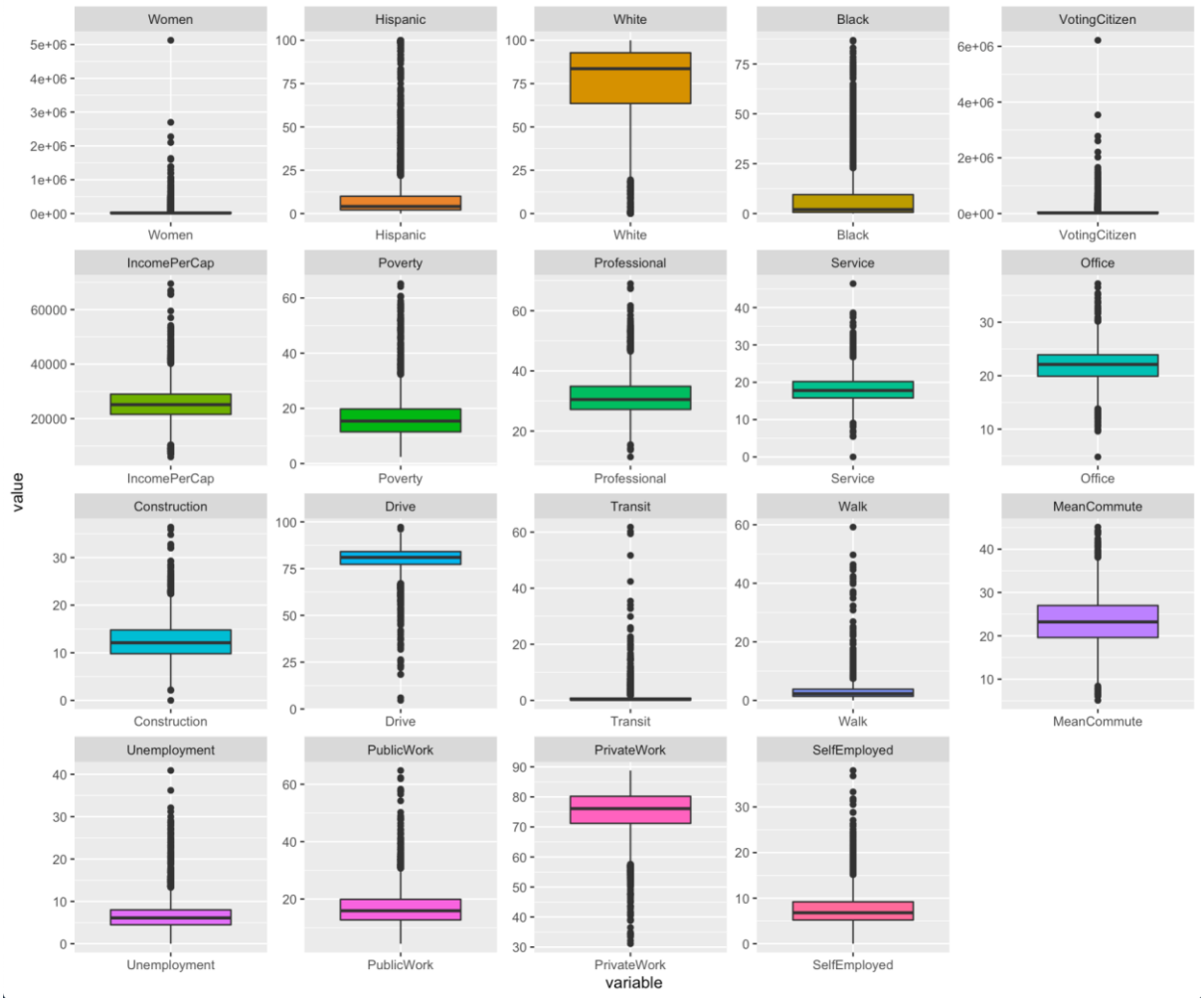


Figure 6. Box-Plots of each variable

A box-plot is a graphical representation of a few statistics: sample mean, quartiles, median, minimum and maximum and, also, it allows to detect outliers.

The outliers are observations that lies an abnormal distance from the other values, sometimes they are wrong measures of the analysis that can distort statistical results, but sometimes they are real values that we shouldn't discard.

In the last figure we can appreciate that the variables with more outliers are Hispanic, Income per capita, Transit and Public Work.

In contrast, the variables with less quantity of outliers are, Women, Poverty, Unemployment and, of course, Mean Commute because, as we have already pointed, this variable has mean values.

The cases with more outliers will be treated individually later on, in order to detect if they can be deleted from the original data or not.

Since, most of the variables have a group of observations that seem to be outliers, there could be some groups that have a different behaviour in an overall view, this will be determined by

implementing some classification tools and multivariate plots that will be shown in the following section and can be used to search for predominant patterns and exceptions.

2.3.3. Multivariate Plots

A parallel coordinates plot shows each observation as a line that connects all the variables, since there are variables with different units, they have been standardized and centered at a value, converting all to the same scale. For the parallel plot we have considered the id variable “region” since it is easier to identify the large amount of counties by area. As we can see, there is one group that highly differs from the others, and which behaviour is at the highest and lowest points of the variables, this is conformed by “Puerto Rico”. Also, there are some big differences in zones based on the race component, for example, the Midwest is conformed of counties that present a high rate of white people, in contrast the South Region, which presents the counties with more black and hispanic people. By looking the relations between transportation, it can be seen that there are counties among the Northeast Region that present the higher rates in the use of public transportation, and, lower rates of the other means. On the contrary, the behavior determined for the West shows that more people walk. Also, we can identify some differences in the type of work, the West concentrates more counties with lower rate of public work, conversely, the Midwest shows counties with higher rates of self employment and more public work. Moreover, the poverty and unemployment rates show some clear patterns for Puerto Rico and the South, as they own the highest rates.

At a first multivariate glance, we can say that there are in fact some patterns and relationships among some variables that will be further studied in the next sections.

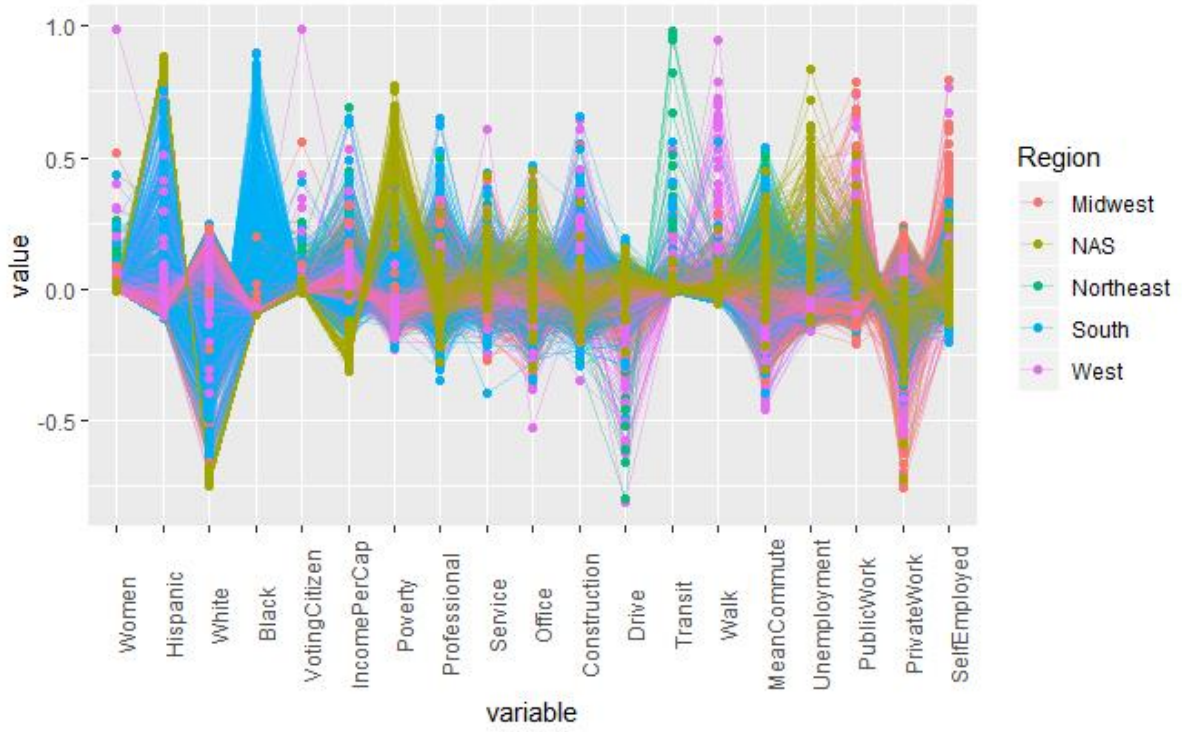


Figure 7. Parallel coordinates plot for all quantitative variables. Region *NAS* refers to Puerto Rico.

An Andrews plot is a graphical data analysis technique for plotting multivariate data. Each multivariate observation conforms one curve, then there are 3220 curves. In order to give more viewable insights, the variable region has been used as previously noted, for dividing the counties. As it can be observed, there are some different behaviours between the counties of different regions and several observations can be distinguished, among all, to be far and might be outliers.

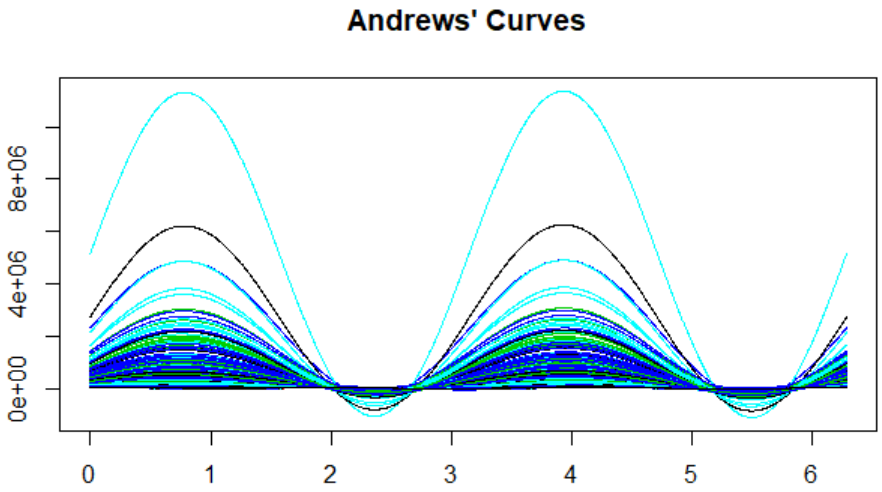


Figure 8. Andres Curves for all quantitative variables

2.3.4. Transformation of Variables

Since there are several variables with asymmetric distributions, as the kernel density plots and the QQ-plots showed, the following transformations have been performed.

Quantitative variables	Variable	Transformation
1	Women	Logarithm
2	Hispanic	Logarithm of variable +1
3	White	Tukey's Ladder of Powers (100 - variable) ^ 0.175
4	Black	Tukey's Ladder of Powers (variable) ^ 0.225
5	VotingCitizen	Logarithm
6	IncomePerCap	NA
7	Poverty	Logarithm
8	Professional	NA
9	Service	NA
10	Office	NA
11	Construction	Logarithm of variable +1
12	Drive	Cubic
13	Transit	Tukey's Ladder of Powers (-1*(1+variable)^(-1.65))
14	Walk	Logarithm of variable +1
15	MeanCommute	NA
16	Unemployment	Logarithm of variable +1
17	PublicWork	Logarithm
18	PrivateWork	Cubic
19	SelfEmployed	Logarithm of variable +1

Table 5. Transformations for quantitative variables

For some variables, there were no evident transformations that fit the data into a normal distribution, for those, we used the Tukey Ladder of Powers, this method allows to find a lambda parameter used in a power transformation as a way of re-expressing the variables. There are several packages in R that facilitate the process, we used the *rcompanion* library, and through the command *transformTukey* the lambda was obtained. If lambda is greater than zero the transformation may be obtained by elevating the variable to the parameter lambda, if the lambda is less than zero, the transformation implies multiplying -1 per the variable elevated to lambda. Moreover, if the variables present high skewness to the right or left some constant is needed. These considerations have been taken in order to obtain the most accurate transformation.

After performing the transformations, it can be observed that the density functions and the QQ-plots show that all treated variables can be considered to be approximately normally distributed.

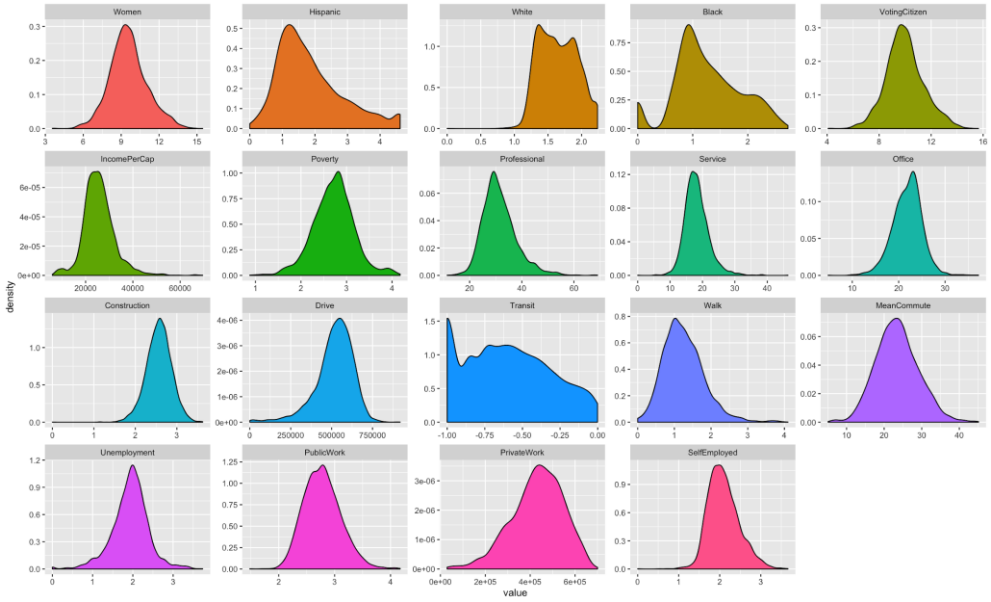


Figure 9. Density Plots after Transformations

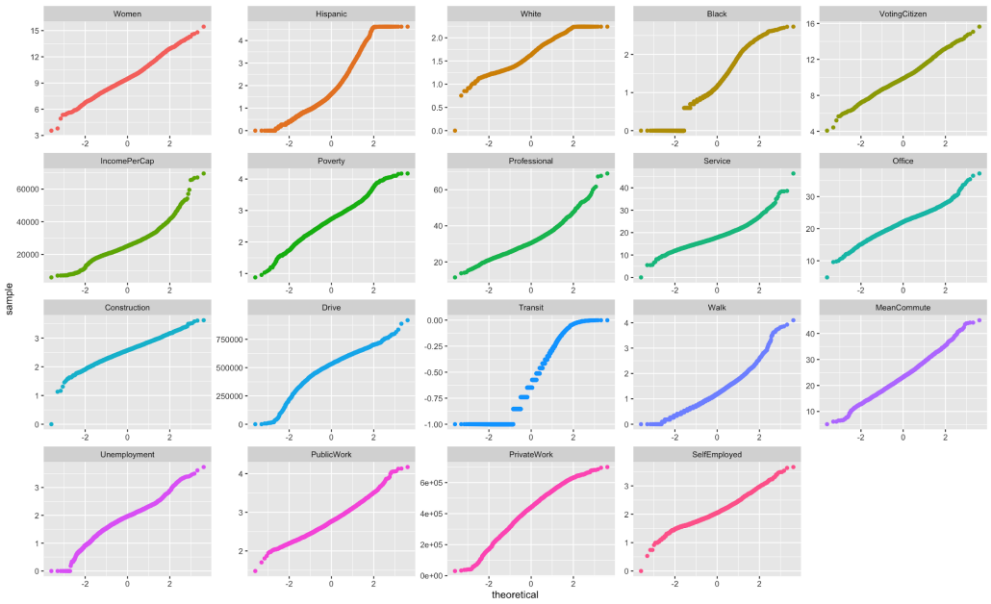


Figure 10. QQ Plots after Transformations

2.4. Outliers

2.4.1. Mahalanobis distance

The Mahalanobis distance is a measure of the distance from a point and a distribution, commonly used for anomaly detection. Its calculation is based on the multiplication of the transpose distance between the vector of the observation and the mean vector times the inverse covariance matrix of independent variables, and multiply again by the distance. The formula is presented below:

$$D_M(x, \mu_x)^2 = (x - \mu_x)' \Sigma_x^{-1} (x - \mu_x)$$

Since the sample mean vector and the sample covariance matrices are largely influenced by outliers, for the calculations, the robust estimators for both metrics have been used, through the calculation of the Minimum Covariance Determinant (MCD) estimator.

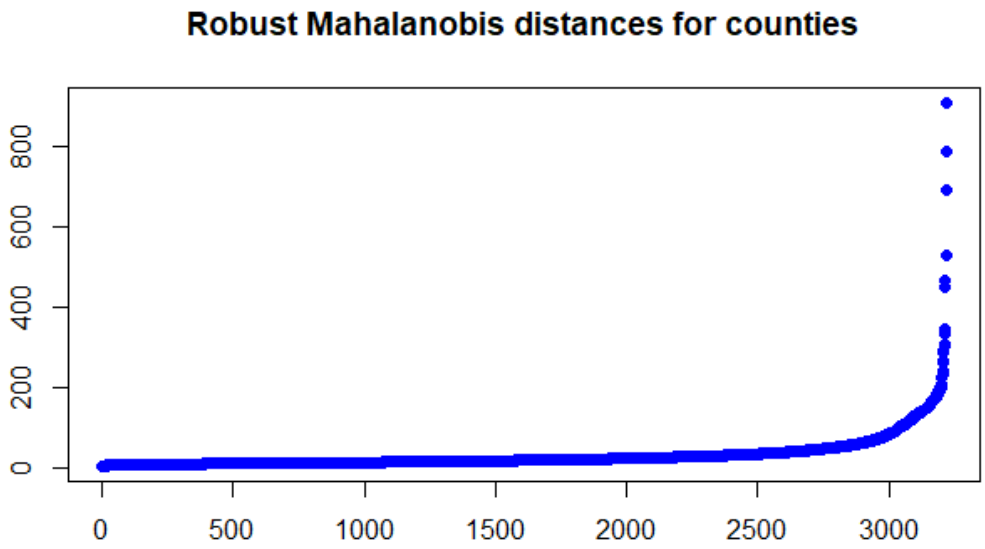


Figure 11. Mahalanobis distance with robust estimators

As it is seen in the resulting graph from the Mahalanobis distance (Figure 11), there are some counties which are far away from the mean in each group. In order to assess a determination of the counties which have a different behavior from the expected one, by assuming Gaussianity, a method based on the False Discovery Rate has been performed.

The final results showed 532 observations that can be considered as outliers. In order to visualize in a macro view, the counties that have been identified as outliers, we have assigned the name of state to which these counties are part of. As we can observe in Table 6, there are some states which have a lot of counties that have an abnormal behaviour. However, these will not be considered as outliers since they are in fact real data and are records that are known to be accurate and true.

It's important to highlight that Puerto Rico, as identified in the previous analysis, has a total behavior far from the center mass, maintaining a total of 78 counties ,from 78, qualified as outliers. Puerto Rico, in an overall basis has a different behaviour that could be understood because of its condition as Free Associated State. Puerto Rico is an unincorporated territory of the United States, with its own legislative body, and the conditions differ in terms of taxation, social benefits and allocation of resources - Puerto Rico follows its own Tax Code since 2011, which differs from the US. These factors lead us to consider the region as an outlier, because its counties can not be treated equally as any other county that belongs to a State. Through this analysis it has been determined to remove Puerto Rico to avoid that its information affects the overall data from the rest of counties.

State	Counties	Counties_out	Percentage_out	Region
Puerto Rico	78	78	1.00	NAS
District of Columbia	1	1	1.00	South
Alaska	29	26	0.90	West
Hawaii	5	4	0.80	West
Nevada	17	9	0.53	West
South Dakota	66	31	0.47	Midwest
North Dakota	53	23	0.43	Midwest
Montana	56	24	0.43	West
New Mexico	33	12	0.36	West
Colorado	64	21	0.33	West
Texas	254	72	0.28	South
Maryland	24	6	0.25	South
Virginia	133	33	0.25	South
Utah	29	7	0.24	West
Nebraska	93	22	0.24	Midwest
Massachusetts	14	3	0.21	Northeast
Mississippi	82	17	0.21	South
Idaho	44	9	0.20	West
Arizona	15	3	0.20	West
California	58	10	0.17	West
Washington	39	6	0.15	West
Florida	67	10	0.15	South
Wyoming	23	3	0.13	West
West Virginia	55	7	0.13	South
Georgia	159	20	0.13	South
Connecticut	8	1	0.13	Northeast
New York	62	7	0.11	Northeast
Kansas	105	11	0.10	Midwest
New Jersey	21	2	0.10	Northeast
Kentucky	120	9	0.08	South
Louisiana	64	4	0.06	South
Arkansas	75	4	0.05	South
Missouri	115	6	0.05	Midwest
Alabama	67	3	0.04	South
North Carolina	100	4	0.04	South
Illinois	102	4	0.04	Midwest
Michigan	83	3	0.04	Midwest
Indiana	92	3	0.03	Midwest
Tennessee	95	3	0.03	South
Pennsylvania	67	2	0.03	Northeast
Wisconsin	72	2	0.03	Midwest
Oregon	36	1	0.03	West
Minnesota	87	2	0.02	Midwest
Ohio	88	2	0.02	Midwest
South Carolina	46	1	0.02	South
Oklahoma	77	1	0.01	South
Vermont	14	0	0.00	Northeast
Rhode Island	5	0	0.00	Northeast
New Hampshire	10	0	0.00	Northeast
Maine	16	0	0.00	Northeast
Iowa	99	0	0.00	Midwest
Delaware	3	0	0.00	South

Table 6. Counties quantified as outliers per State

2.5. Standardization of Variables

Standardization is considered as the process of putting different variables on the same scale, this can be achieved through the default version of the command in R *scale***.** The standardization of variables resulting on each variable to have mean zero and standard deviation one, has been done after the removal of outliers, and so it can be used for performing the classification tools that will be presented next in this project.

2.6. Correlation of the Variables

While both, covariance and correlation, measure the relationship between two variables, covariance indicates the direction of the linear relationship between the variables which can take any value between -∞ and +∞. In contrast, correlation measures the strength and the direction of said linear relationship, where all values are standardized, such that its range remains between -1 and +1. The correlation and covariance matrices can be found in figure 12 and table 7.

2.6.1. Correlation and Covariance Matrix

On one hand, we can see that some variables depend on others and are highly correlated to each other. This is the case between Income Per Capita and Poverty, Income Per Capita and Professional, Poverty and Unemployment, Drive and Walk, Public and Private work, etc. with correlation rates ranging between ±0.6 and ±0.84. These relationships suggest that we may need to select some variables that can contribute better our analysis, while discarding others to avoid extra noise in our data.

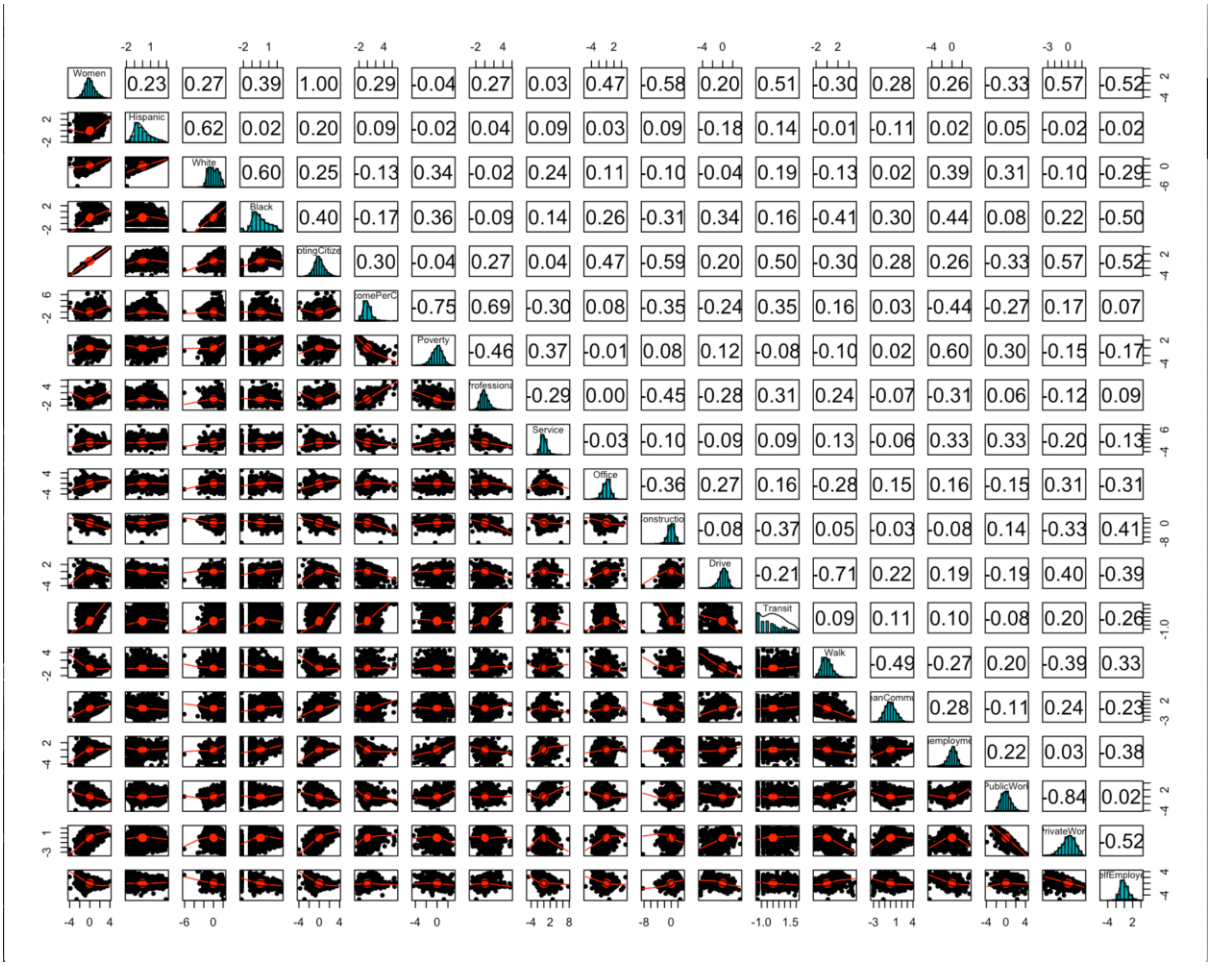


Figure 12. Matrix Correlation and Density Plots

Figure 12 shows the matrix plot of the distributions of each variable by histogram graphs in the diagonal. In addition, this plot represents the correlation between all variables by scatter plots in the lower-triangular part and by their correlation coefficients in the upper-triangular part of the matrix.

	Women	Hispanic	White	Black	VotingCitizen	IncomePerCap	Poverty	Professional	Service	Office
Women	2.248e+00	3.167e-01	1.091e-01	3.414e-01	2.202e+00	2.769e+03	-2.203e-02	2.606e+00	1.513e-01	2.137e+00
Hispanic	3.167e-01	8.592e-01	1.545e-01	1.281e-02	2.770e-01	5.082e+02	-9.444e-03	2.254e-01	3.141e-01	9.345e-02
White	1.091e-01	1.545e-01	7.280e-02	9.325e-02	9.959e-02	-2.192e+02	3.859e-02	-3.450e-02	2.412e-01	8.840e-02
Black	3.414e-01	1.281e-02	9.325e-02	3.334e-01	3.391e-01	-6.078e+02	8.713e-02	-3.504e-01	2.968e-01	4.522e-01
VotingCitizen	2.202e+00	2.770e-01	9.959e-02	3.391e-01	2.165e+00	2.730e+03	-2.379e-02	2.564e+00	2.119e-01	2.116e+00
IncomePerCap	2.769e+03	5.082e+02	-2.192e+02	-6.078e+02	2.730e+03	3.931e+07	-1.973e+03	2.829e+04	-6.891e+03	1.590e+03
Poverty	-2.203e-02	-9.444e-03	3.859e-02	8.713e-02	-2.379e-02	-1.973e+03	1.749e-01	-1.262e+00	5.746e-01	-1.043e-02
Professional	2.606e+00	2.254e-01	-3.450e-02	-3.504e-01	2.564e+00	2.829e+04	-1.262e+00	4.277e+01	-6.892e+00	5.339e-02
Service	1.513e-01	3.141e-01	2.412e-01	2.968e-01	2.119e-01	-6.891e+03	5.746e-01	-6.892e+00	1.359e+01	-3.103e-01
Office	2.137e+00	9.345e-02	8.840e-02	4.522e-01	2.116e+00	1.590e+03	-1.043e-02	5.339e-02	-3.103e-01	9.297e+00
Construction	-2.658e-01	2.575e-02	-8.542e-03	-5.497e-02	-2.646e-01	-6.717e+02	1.060e-02	-9.133e-01	-1.105e-01	-3.332e-01
Drive	3.415e+04	-1.887e+04	-1.220e+03	2.247e+04	3.426e+04	-1.736e+08	5.566e+03	-2.140e+05	-4.023e+04	9.650e+04
Transit	2.196e-01	3.800e-02	1.505e-02	2.627e-02	2.152e-01	6.272e+02	-9.918e-03	5.919e-01	9.815e-02	1.380e-01
Walk	-2.525e-01	-6.663e-03	-1.982e-02	-1.329e-01	-2.449e-01	5.518e+02	-2.361e-02	8.666e-01	2.733e-01	-4.731e-01
MeanCommute	2.326e+00	-5.841e-01	2.289e-02	9.918e-01	2.331e+00	1.076e+03	3.910e-02	-2.722e+00	-1.223e+00	2.611e+00
Unemployment	1.656e-01	9.633e-03	4.453e-02	1.096e-01	1.631e-01	-1.173e+03	1.068e-01	-8.802e-01	5.268e-01	2.098e-01
PublicWork	-1.628e-01	1.655e-02	2.762e-02	1.444e-02	-1.574e-01	-5.500e+02	4.099e-02	1.262e-01	4.014e-01	-1.523e-01
PrivateWork	9.781e+04	-2.001e+03	-3.152e+03	1.424e+04	9.538e+04	1.241e+08	-7.188e+03	-8.963e+04	-8.623e+04	1.076e+05
SelfEmployed	-2.955e-01	-5.907e-03	-2.958e-02	-1.092e-01	-2.902e-01	1.733e+02	-2.741e-02	2.180e-01	-1.862e-01	-3.588e-01

	Construction	Drive	Transit	Walk	MeanCommute	Unemployment	PublicWork	PrivateWork	SelfEmployed
Women	-2.658e-01	3.415e+04	2.196e-01	-2.525e-01	2.326e+00	1.656e-01	-1.628e-01	9.781e+04	-2.955e-01
Hispanic	2.575e-02	-1.887e+04	3.800e-02	-6.663e-03	-5.841e-01	9.633e-03	1.655e-02	-2.001e+03	-5.907e-03
White	-8.542e-03	-1.220e+03	1.505e-02	-1.982e-02	2.289e-02	4.453e-02	2.762e-02	-3.152e+03	-2.958e-02
Black	-5.497e-02	2.247e+04	2.627e-02	-1.329e-01	9.918e-01	1.096e-01	1.444e-02	1.424e+04	-1.092e-01
VotingCitizen	-2.646e-01	3.426e+04	2.152e-01	-2.449e-01	2.331e+00	1.631e-01	-1.574e-01	9.538e+04	-2.902e-01
IncomePerCap	-6.717e+02	-1.736e+08	6.272e+02	5.518e+02	1.076e+03	-1.173e+03	-5.500e+02	1.241e+08	1.733e+02
Poverty	1.060e-02	5.566e+03	-9.918e-03	-2.361e-02	3.910e-02	1.068e-01	4.099e-02	-7.188e+03	-2.741e-02
Professional	-9.133e-01	-2.140e+05	5.919e-01	8.666e-01	-2.722e+00	-8.802e-01	1.262e-01	-8.963e+04	2.180e-01
Service	-1.105e-01	-4.023e+04	9.815e-02	2.733e-01	-1.223e+00	5.268e-01	4.014e-01	-8.623e+04	-1.862e-01
Office	-3.332e-01	9.650e+04	1.380e-01	-4.731e-01	2.611e+00	2.098e-01	-1.523e-01	1.076e+05	-3.588e-01
Construction	9.431e-02	-2.876e+03	-3.284e-02	8.959e-03	-4.354e-02	-1.083e-02	1.425e-02	-1.147e+04	4.821e-02
Drive	-2.876e+03	1.332e+10	-6.922e+03	-4.630e+04	1.420e+05	9.308e+03	-7.356e+03	5.295e+09	-1.695e+04
Transit	-3.284e-02	-6.922e+03	8.388e-02	1.413e-02	1.786e-01	1.251e-02	-7.214e-03	6.632e+03	-2.884e-02
Walk	8.959e-03	-4.630e+04	1.413e-02	3.160e-01	-1.544e+00	-6.509e-02	3.763e-02	-2.495e+04	7.158e-02
MeanCommute	-4.354e-02	1.420e+05	1.786e-01	-1.544e+00	3.179e+01	6.790e-01	-2.003e-01	1.522e+05	-4.939e-01
Unemployment	-1.083e-02	9.308e+03	1.251e-02	-6.509e-02	6.790e-01	1.835e-01	3.165e-02	1.673e+03	-6.131e-02
PublicWork	1.425e-02	-7.356e+03	-7.214e-03	3.763e-02	-2.003e-01	3.165e-02	1.082e-01	-3.160e+04	2.177e-03
PrivateWork	-1.147e+04	5.295e+09	6.632e+03	-2.495e+04	1.522e+05	1.673e+03	-3.160e+04	1.305e+10	-2.274e+04
SelfEmployed	4.821e-02	-1.695e+04	-2.884e-02	7.158e-02	-4.939e-01	-6.131e-02	2.177e-03	-2.274e+04	1.451e-01

Table 7. Covariance Matrix

2.6.2. Eigenvalues

To better visualize and understand the previously established relationships between the variables, the correlation matrix has been included below. From this plot, it can be observed that there are actually groups of correlated variables that may suggest a factor structure. In order to further investigate these structures, the eigenvalues have been computed and graphed along with the explained variation for which each variable is responsible for.

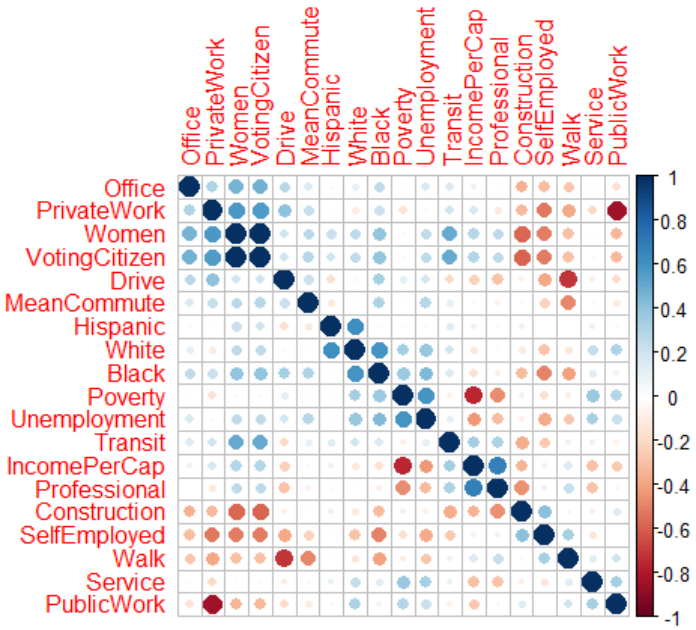


Figure 14. Correlation Matrix of ordered terms

From the table below, it can be observed that the first variable accounts for 26% of the variations, while all other variables have progressively lower contributions. The direction of the first 5 eigenvectors play a dominant role, as roughly 73% of the variability is explained by them. To further uncover some structure within the data and the relationships between variables, we will review Principal Component and Factor Analysis in sections to come.

	eigenvalue	variance.percent	cumulative.variance.percent
1	4.953638	26.071780	26.07
2	3.636830	19.141211	45.21
3	2.555775	13.451447	58.66
4	1.337813	7.041122	65.71
5	1.223183	6.437806	72.14
6	0.975372	5.133536	77.28
7	0.773795	4.072604	81.35
8	0.682004	3.589496	84.94
9	0.557385	2.933603	87.87
10	0.543019	2.857996	90.73
11	0.474193	2.495752	93.23
12	0.348142	1.832329	95.06
13	0.315328	1.659620	96.72
14	0.188594	0.992598	97.71
15	0.175291	0.922582	98.63
16	0.138295	0.727870	99.36
17	0.105042	0.552854	99.91
18	0.015129	0.079625	99.99
19	0.001172	0.006169	100.00

Table 8. This table portrays the eigenvalues corresponding to the quantitative variables in the dataset, along with the progression of percentage values given the variation for which each variable is responsible. The last column represents the cumulative progression of variation.

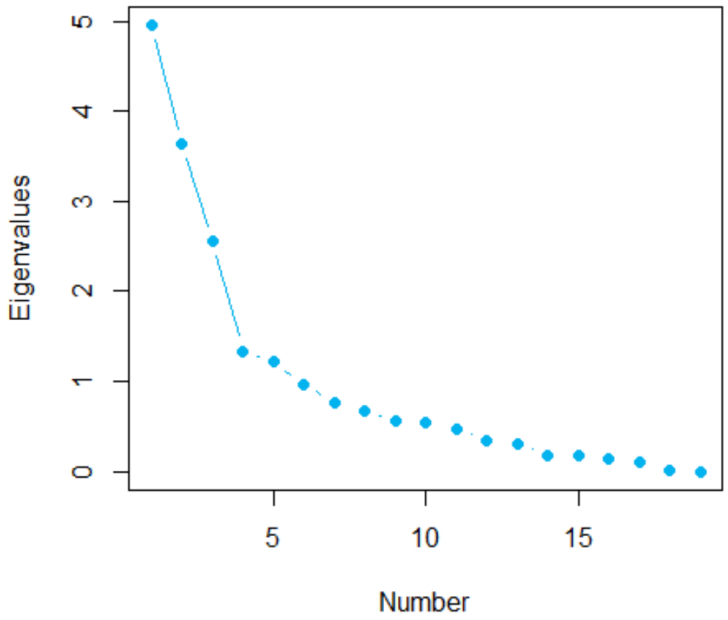


Figure 15. Eigenvalues

3. PCA

In cases where many variables are present, you cannot easily plot the data in its raw format, making it difficult to get a sense of the trends present within. PCA (Principal Component Analysis) allows us to see the overall "shape" of the data, identifying which samples are similar to one another and which are very different. This enables the identification of groups of samples that are similar and determine which variables make one group differ from another.

This statistical procedure uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

Then, the eigenvector is a direction, such as "vertical" or "45 degrees", while an eigenvalue is a number telling you how much variance there is in the data in that direction. The eigenvector with the highest eigenvalue is, therefore, the first principal component.

If your initial variables are strongly correlated with one another, you will be able to approximate most of the complexity in your dataset with just a few principal components. As you add more principal components, you summarize more and more of the original dataset. Adding additional components makes your estimate of the total dataset more accurate, but also more unwieldy.

As we have seen above in the section 2.6.1 our variables are highly correlated in many cases, so we can assume with some certainty that the application of a PCA transformation here can help us to better understand our data.

The PCA only can be used in quantitative variables, we have excluded the variables of "County" and "State.The next step would be to standardize and center the data, which was already done previously in section 2.5.

We have 3,142 records and 19 variables ready to be implemented with the prcomp() function in R.

We have then, the 19 PCAs with their corresponding characteristics:

```
Importance of components:
                          PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10  PC11   PC12   PC13    PC14
Standard deviation     2.226 1.907 1.599 1.1566 1.1060 0.9876 0.8797 0.8258 0.7466 0.7369 0.689 0.5900 0.5615 0.43427
Proportion of Variance 0.261 0.191 0.135 0.0704 0.0644 0.0513 0.0407 0.0359 0.0293 0.0286 0.025 0.0183 0.0166 0.00993
Cumulative Proportion  0.261 0.452 0.587 0.6571 0.7214 0.7728 0.8135 0.8494 0.8787 0.9073 0.932 0.9506 0.9672 0.97711
                          PC15    PC16    PC17   PC18    PC19
Standard deviation     0.41868 0.37188 0.32410 0.1230 0.03424
Proportion of Variance 0.00923 0.00728 0.00553 0.0008 0.00006
Cumulative Proportion  0.98633 0.99361 0.99914 0.9999 1.00000
```

Figure 15. Summary of the 19 PCAs.

Let's visualize now some graphs about the first PCAs:

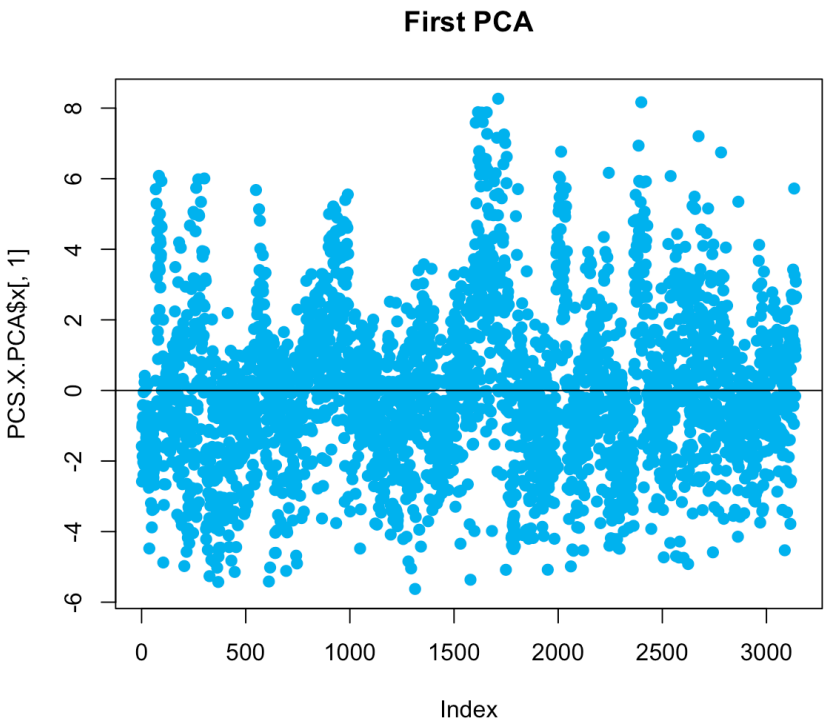


Figure 16. First PCA distribution.

We can observe that already the first PCA can separate the data of the sides more far away from 0 in two groups, nevertheless, there is still a certain continuity throughout the points.

Let's visualize the first two PCAs as we can plot them in two dimensions:

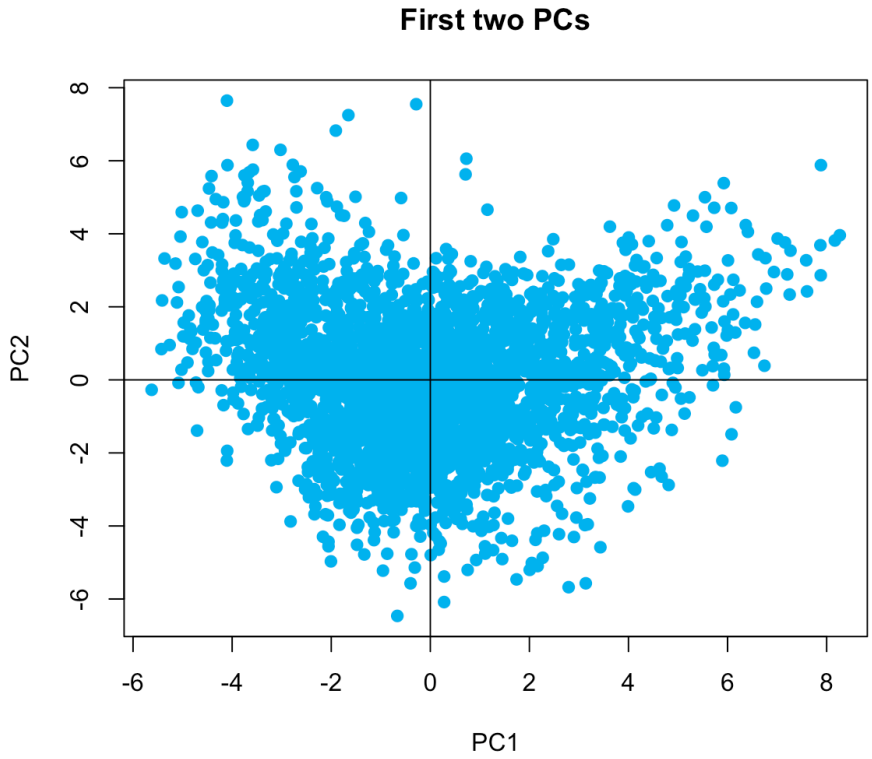


Figure 16. First two PCAs.

We can see now that there is more distinction between the groups formed by the quadrants. The data from the positive right quadrant, that is far away form the x axis, will have many more different characteristics than the data of the positive left quadrant, that is far away form the x axis, and it happens the same with the y axis.

With the function *ggbiplot* we can now plot the distribution of our data into the first two PCAs showing how the initial variables map onto it and simultaneously revealing how each one contribute to each principal component.

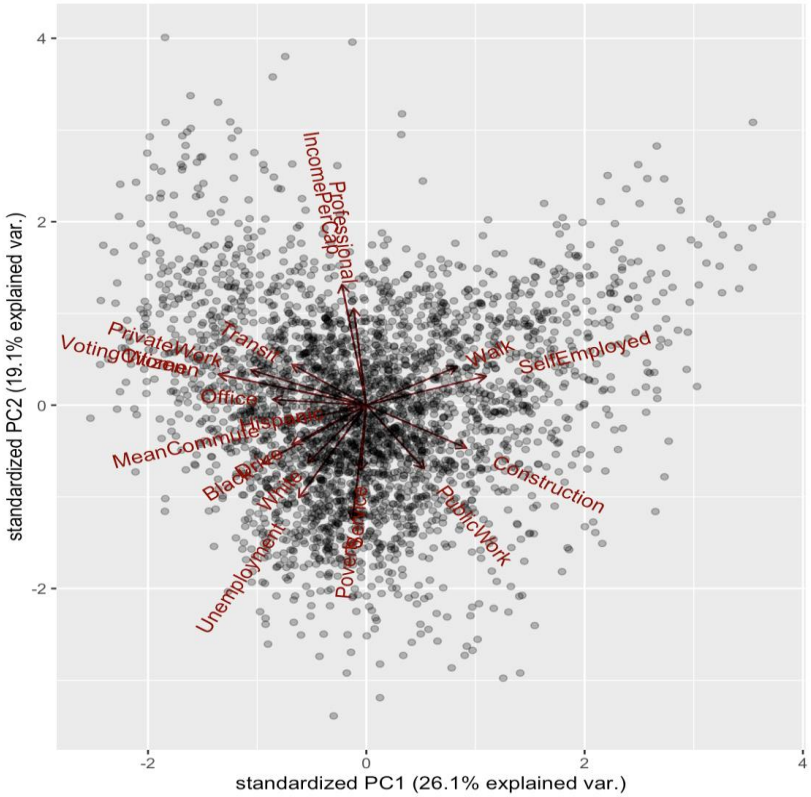


Figure 17. First and Second PCA distributions with initial variables.

Then, the initial variables that contribute the most to the first PCA are Self Employed, Construction, Women, Voting Citizen, Private Work and Office, on the contrary, the ones that contribute the most to the second PCA are Professional, Income per Capita, Poverty, and Unemployment. The most important variables for both groups can be identified clearly in the next plot:

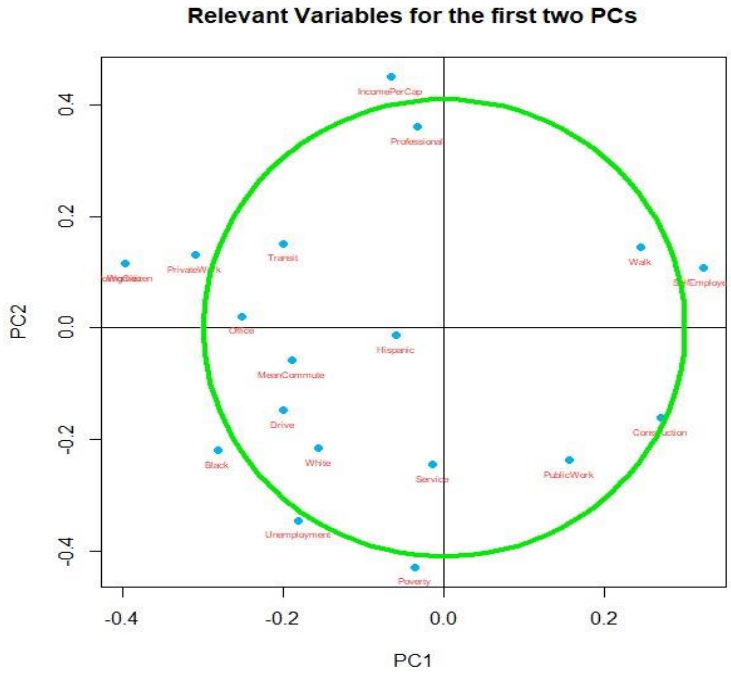


*Figure 18.*Relevant initial variables for both PCAs.

However, the first and second PCA only represent 45.2% of the variability of the data which is not too high to only take them into account. Let's then obtain how many PCAs are needed to better explain all the data.

For doing so, we need to check the percentages of the explained variances per all the dimensions (PCAs) which are the eigenvalues:

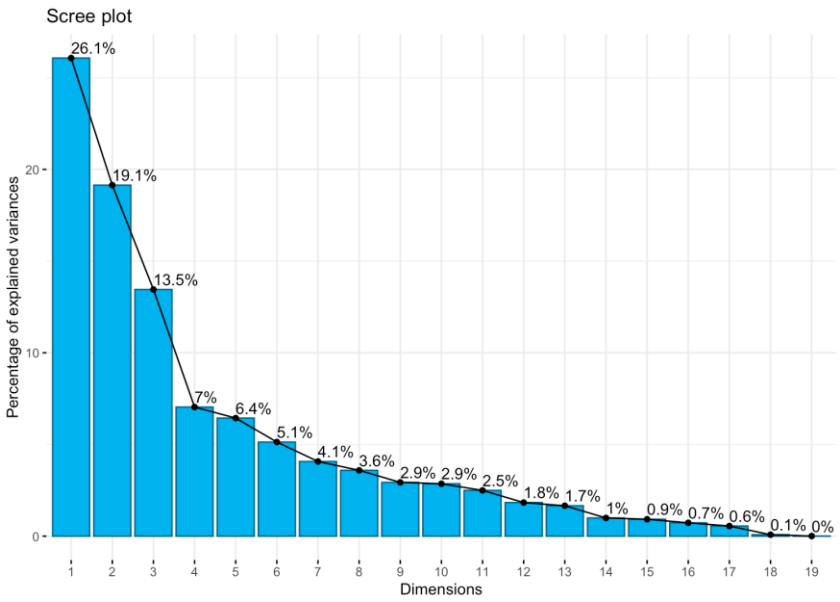


Figure 19. Variabilities Percentages for PCAs

In addition, the detailed list as previously showcased in section 2.6.3:

Eigenvalues	Variance Percent	Cumulative Variance Percent
4.9536	26.0718	26.0700
3.6368	19.1412	45.2100
2.5558	13.4514	58.6600
1.3378	7.0411	65.7100
1.2232	6.4378	72.1400
0.9754	5.1335	77.2800
0.7738	4.0726	81.3500
0.6820	3.5895	84.9400
0.5574	2.9336	87.8700
0.5430	2.8580	90.7300
0.4742	2.4958	93.2300
0.3481	1.8323	95.0600
0.3153	1.6596	96.7200
0.1886	0.9926	97.7100
0.1753	0.9226	98.6300
0.1383	0.7279	99.3600
0.1050	0.5529	99.9100
0.0151	0.0796	99.9900
0.0012	0.0062	100.0000

Table 10. Eigenvalues and its corresponding explained variance percentages.

As we can see in the table 10, the first 5 PCAs explain more than 70% of the variability, then, the dimension of the data set may be reduced from 19 to 5 that represent the number of new variables (*Principal Components*) keeping more than 70% of the information inside.

The eigenvectors of the sample covariance matrix of X are given in rotation and show how much the initial variables influence to each PCA, the higher the absolute value is, the more they contribute to the PCA. The details are shown in the next table:

	PC1	PC2	PC3	PC4	PCA5
Women	-0.3978	0.1149	-0.1155	0.0143	-0.10275
Hispanic	-0.0600	-0.0134	-0.2913	0.6902	-0.14312
White	-0.1575	-0.2159	-0.3704	0.4039	0.104604
Black	-0.2815	-0.2203	-0.0853	0.0083	0.251326
VotingCitizen	-0.3978	0.1154	-0.1122	-0.0097	-0.09791
IncomePerCap	-0.0657	0.4507	-0.1159	0.0533	0.179797
Poverty	-0.0369	-0.4293	-0.0773	-0.1430	-0.16307
Professional	-0.0341	0.3616	-0.2635	-0.1135	0.358493
Service	-0.0146	-0.2442	-0.2527	-0.2359	-0.30028
Office	-0.2519	0.0211	0.0348	-0.0559	0.023092
Construction	0.2696	-0.1606	0.1656	0.3748	-0.02214
Drive	-0.2005	-0.1483	0.3765	0.0205	0.196369
Transit	-0.1998	0.1517	-0.2887	-0.1138	-0.13702
Walk	0.2446	0.1443	-0.3089	-0.2137	-0.32568
MeanCommute	-0.1889	-0.0587	0.1594	-0.0015	0.377228
Unemployment	-0.1813	-0.3454	-0.0974	-0.1355	-0.00574
PublicWork	0.1550	-0.2377	-0.3462	-0.1539	0.415295
PrivateWork	-0.3102	0.1313	0.2969	0.0799	-0.35673
SelfEmployed	0.3225	0.1087	0.0067	0.1317	0.034369

Table 11. Weights of initial variables for each first five PCAs.

Then, the variables that are highlighted in orange are the ones that contribute the most to their corresponding PC. We can also visualize it with the next graph:

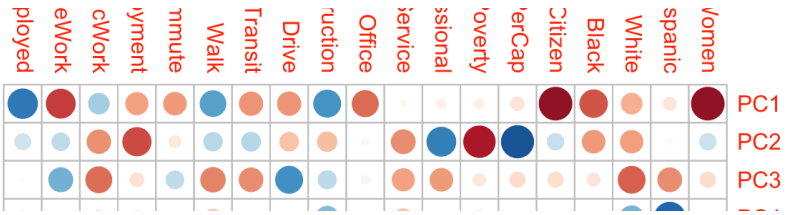


Figure 20. Correlations between initial variables and first five PCAs

After reviewing the results from the important variables of each of the PCAs, we can identify that there are common characteristics between these groups at least for the first and second PCAs. The variables given by the first PCA can be related to “Configuration of population”, as we can see here the variables Women, Black and Voting Citizens, which can be related to the adult workforce, have an important role, also the sector of their work. The second PCA can be identified as “Wealth” since the variables that mostly contribute to the group variability are Poverty and Unemployment in one side and, Income per capita and two types of labor that are very related to economic progress which are Professional, which consider the people who are mostly highly educated and in administrative roles. These two PCAs explain 45.2% of the total variability of the data, which as said before, is not two high but in macro terms it is really useful that these dissemination allows to know that almost have of the variability of the data will be given by fewer variables and if we add some more variables we could easily explained the majority of it.

Unfortunately, we can't plot the five PCs as we cannot visualize five dimensions, but we can see how PCAs work in group classifications for qualitative variables like, for instance, "Region" in the next figure.

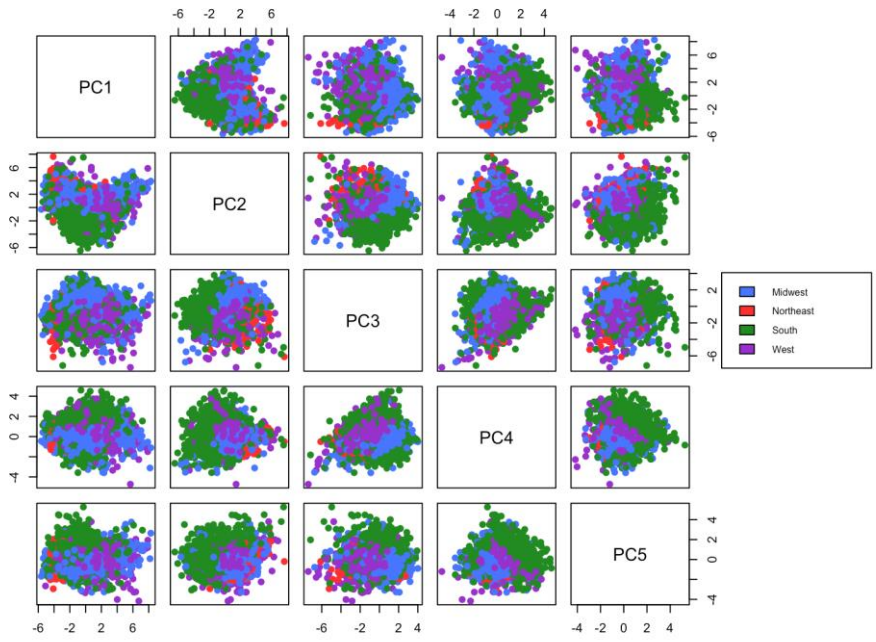


Figure 21. Scores of five PCAs for the four groups of Regions

The pattern that was identified in Figure 16, now can be explained. As seen in the plots, PCA shows that there is a different behaviour for the South Region, this is more evident by relating the second component, with the rest. As seen, there is a marked difference between this Region and what can be called as the non-South, also evident when relating the fourth PCA, which if we review gets its importance from the variable Hispanic. Consequently, can be understood that wealth and the percentage of hispanic people can be attributes that characterize and divide implicity to US counties.

4. Factor Analysis

The aim of factor analysis (FA) is to explain the outcome of p variables in the data matrix X using fewer variables, the so-called factors. Ideally, all the information in X can be produced by a smaller number of factors. These factors are interpreted as latent (unobserved) common characteristics of the observed $x \in \mathbb{R}^p$. Although principal component analysis and factor analysis might be related, they are quite different in nature. PCs are linear transformations of X arranged in decreasing order of variance and used to reduce the dimension of the data set, whereas in factor analysis, we try to model the variations of X using a linear transformation of a fixed, limited number of latent factors (Härdle, Simar 2019). The factor model explains not only the underlying structure among the latent variables, but also among the observed variables and the relationship between them.

In FA, the number of factors are unknown and smaller than the number of observed variables. The factor model is given such that X can be explained as a function of the mean of the multivariate random variables, the loading matrix (L) of unknown constants, the factors and the errors, where the factors (f) and errors are uncorrelated.

$$x = \mu_x + Lf + \epsilon$$

As in PCA, the covariance matrix plays a vital role in the factor model given the relationship between it and the loading matrix,

$$\Sigma_x = E\left[(x - \mu_x)(x - \mu_x)'\right] = LL' + \Sigma_\epsilon ,$$

where the loading matrix is the covariance matrix between the multivariate random variables and the latent factors.

$$Cov[x, f] = E\left[(x - \mu_x) f'\right] = E\left[(Lf + \epsilon) f'\right] = L$$

The estimation of the factor model can be done in several different ways, through the *Maximum Likelihood* method, *Principal Component Factor* method, and *Principal Factor* method. We will explore these methods in the following sections.

4.1.1. Principal Component Factor Analysis

In order to identify the latent factors of our data using PCFA, we must estimate our loading matrix, the factor scores, and the diagonal covariance matrix. The first few steps have already been performed in previous sections. First we scale the data, then we estimate the covariance using the

sample correlation matrix of X. Retrieving the eigenvalues, we discard the smaller ones and, like in PCA, we select the number of factors (r), which we will keep at 5.

We now proceed to the estimation of our loading matrix (M), where we will use the varimax criterion to better interpret the results. The values of the loading matrix show the estimated correlation between each of the 19 variables and the five factors. As it can be observed below in table 12, given the highlighted terms, the first factor seems to be an index of social determinants, as most related to the variables *Women, Voting Citizen, Construction, Transit,* and *Self-Employed*, this factor can have a really similar interpretation as the called "Configuration of the population" in the PCA . The second factor seems to be an index of the overall economic well-being of the regions or almost equally to "Wealth" PCA. The third factor may be attributed to socioeconomic index, as it includes African Americans along with the variables measuring the mean time spent commuting to work, and whether is by car (Drive) or walking (Walk). The fourth factor includes the white population, which is the racial majority and Hispanics, which are the largest racial minority, so it may be attributed to ethnicity. Lastly, it's known that those employed in public offices earn significantly less than those in the private sector and that the private sector is much larger than the public, generating jobs at a much faster rate; therefore, we may see the fifth factor as an index of business organization among the regions.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Women	-0.837671	0.03904	0.22907	0.23042	-0.265416
Hispanic	-0.027863	0.05234	-0.13144	0.93348	-0.073686
White	-0.239257	-0.28746	0.14345	0.78959	0.301142
Black	-0.408504	-0.33563	0.54537	0.2286	0.189349
VotingCitizen	-0.844898	0.03592	0.23033	0.20382	-0.257781
IncomePerCap	-0.383427	0.81866	-0.13935	0.03138	-0.007401
Poverty	0.028234	-0.84651	0.08285	0.05728	0.156768
Professional	-0.431237	0.69503	-0.15507	-0.02158	0.373532
Service	-0.197069	-0.63586	-0.29703	0.04409	0.180447
Office	-0.448857	-0.02843	0.30896	0.0107	-0.161412
Construction	0.815733	-0.09269	-0.0305	0.18646	-0.05177
Drive	0.006214	-0.1647	0.75214	-0.15128	-0.271014
Transit	-0.681848	0.08146	-0.20368	0.14681	0.015419
Walk	0.026674	0.06772	-0.86751	-0.09891	0.197795
MeanCommute	-0.155676	0.03793	0.63096	-0.05057	0.043549
Unemployment	-0.261883	-0.67301	0.29173	0.10458	0.166871
PublicWork	0.140124	-0.2561	-0.07256	0.08679	0.880531
PrivateWork	-0.414658	0.03986	0.29485	-0.03649	-0.816739
SelfEmployed	0.570619	0.31354	-0.3756	-0.03311	0.129467

Table 12. Estimation of the matrix M using the varimax rotation.

Now let us work "backwards" and find the variables which are better explained by the factors. We do this through the computation of the communalities. The variables better explained by their respective factors are *Private Work, Hispanic,* and *Women* - corresponding to the factors indexing business organization, ethnicity and social determinants. Then, estimating the covariance matrix of the errors, we can compute the uniqueness, which will help us identify which variables are poorly explained by the factors. From the results below, the variables worst explained by the factors are *Office, Mean Commute*, and *Transit* - which is actually not significant within any of the factor structures.

PrivateWork	Hispanic	Women	VotingCitizen	White	PublicWork	IncomePerCap	Professional	Walk	Poverty
0.9289	0.8976	0.8792	0.8762	0.8746	0.8734	0.8377	0.8331	0.8068	0.7521
Construction	Drive	Black	Unemployment	SelfEmployed	Service	Transit	MeanCommute	Office	
0.7124	0.6892	0.6651	0.6454	0.5828	0.5659	0.5348	0.4282	0.3239	

Office	MeanCommute	Transit	Service	SelfEmployed	Unemployment	Black	Drive	Construction	Poverty
0.67610	0.57176	0.46517	0.43412	0.41715	0.35459	0.33493	0.31079	0.28761	0.24790
Walk	Professional	IncomePerCap	PublicWork	White	VotingCitizen	Women	Hispanic	PrivateWork	
0.19322	0.16694	0.16232	0.12664	0.12541	0.12381	0.12077	0.10239	0.07114	

Table 13. Communalities and Uniqueness, refer to the relationship between the variables and the factors as they show how well or poorly the variables are explained by them.

We can already see the difference between the results from PCFA and those from the PCA previously conducted. In the PCA, variables were found repeatedly throughout the 5 components, which isn't the case through PCFA; this is not surprising given the distinct goals of each analysis. PCA focuses on reduction of dimensionality through the components, while Factor Analysis focuses on better explaining the data and understanding the existing relationships through the factors.

Checking the correlation between factors, it's confirmed that there is none.

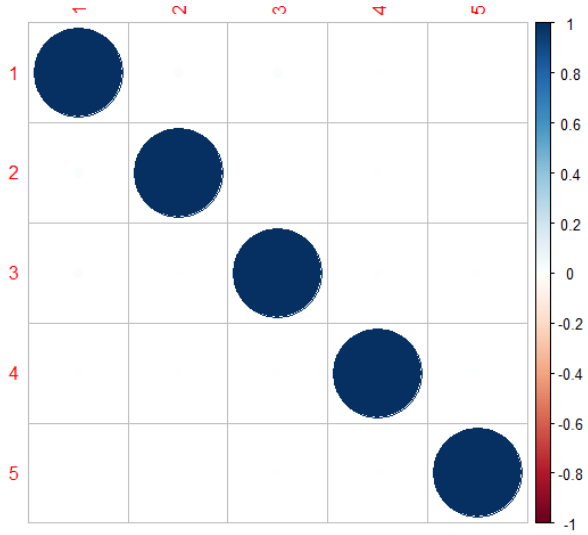


Figure 22. Correlation plot of estimated factor scores.

If we analyze the plot of the correlation among residuals, given the factor scores and the loading matrix , we are able to identify the conglomerates and confirm the previously established factors and their associated variables. However, it also shows that there are some correlations that the model was not able to explain.

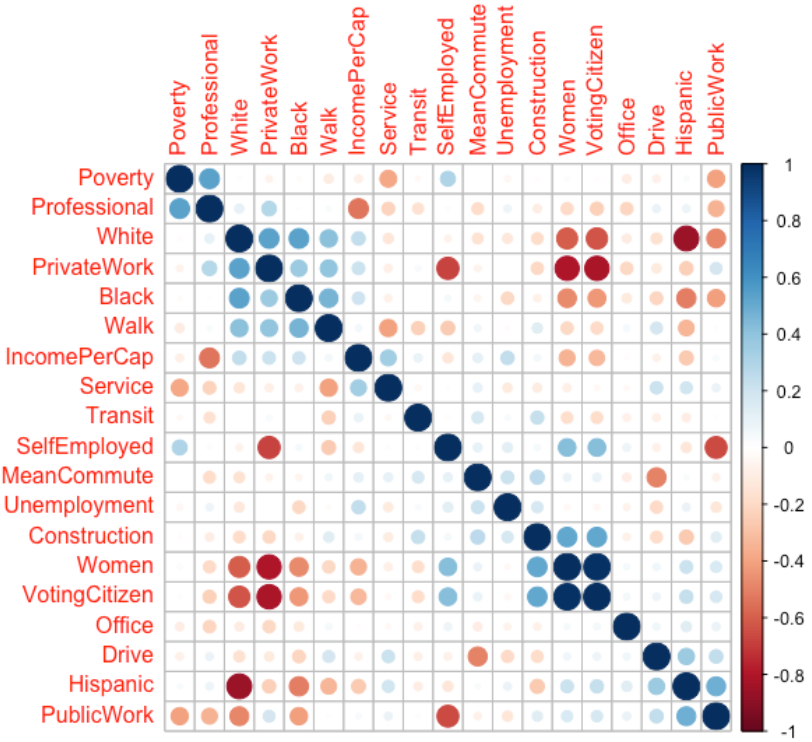


Figure 23. Correlation plot of the estimated residuals given the factor scores and the loading matrix.

4.1.2. Principal Factor Analysis

Through this method, we complete the same first steps as with PCFA. However, to calculate our loading matrix, we must first obtain a matrix that expresses the difference between the correlation matrix of X and the covariance matrix of the errors, which yield the spectral decomposition. Then retrieving the eigenvectors and values, we can estimate the loading matrix (M), using the varimax criterion again for interpretation. The composition of the factors through PFA, yield very similar results to the previous method. As it can be confirmed with the data in table 14, this method has preserved the same variables we observed in PCFA, with the exception of *Mean Commute*, which is not part of any of the factors. The first 2 factors have maintained the exact same variables as the previous method, as well as the fourth factor. Now, the third factor here is composed of public and private work, which was the fifth factor in PCFA, so these two factors have been interchanged. Through this method, the relation between *Self Employed* and the first factor is not as close to 0.6, so we may consider excluding it - likewise with Black.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Women	-0.84332	-0.02135	-0.24876	-0.20698	0.25141
Hispanic	-0.03977	-0.04189	-0.06346	-0.91781	-0.12175
White	-0.24977	0.30833	0.30557	-0.74759	0.14305
Black	-0.39004	0.33634	0.16624	-0.1906	0.50862
VotingCitizen	-0.84941	-0.01852	-0.24161	-0.18131	0.25238
IncomePerCap	-0.39348	-0.79594	-0.0176	-0.03754	-0.15429
Poverty	0.03724	0.81998	0.145	-0.04306	0.09202
Professional	-0.44551	-0.66148	0.3456	0.01882	-0.18697
Service	-0.14822	0.55597	0.16493	-0.06445	-0.21369
Office	-0.38766	0.02073	-0.12398	-0.02507	0.29373
Construction	0.76656	0.08947	-0.0431	-0.14715	-0.06104
Drive	0.01778	0.11875	-0.19461	0.13268	0.75169
Transit	-0.617	-0.05777	-0.01083	-0.12864	-0.15666
Walk	0.01624	-0.04096	0.14286	0.08673	-0.87464
MeanCommute	-0.15945	0.01575	-0.03159	0.03753	0.47556
Unemployment	-0.238	0.64081	0.13633	-0.09216	0.27798
PublicWork	0.13498	0.2681	0.85596	-0.0839	-0.10925
PrivateWork	-0.40424	-0.05241	-0.80088	0.03919	0.34558
SelfEmployed	0.51612	-0.27819	0.10034	0.03998	-0.39324

Table 14. Estimation of the matrix $\widehat{M}$, using varimax criterion.

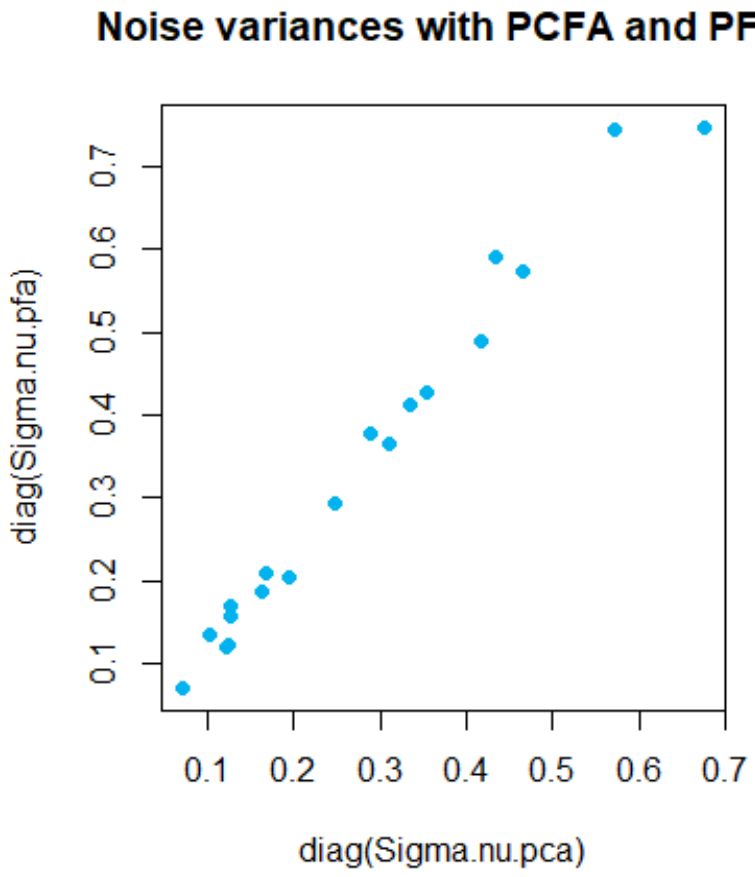


Figure 24. Comparison between PFA and PCFA estimates of the covariance matrix of the errors.

Comparing the PFA and the PCFA graphically, we can observe some minor differences. overall the communality values are higher under PCFA, as well as the uniqueness; however, we can observe the best and worst explained variables by the factors have changed. The variables best explained by the factors are Private Work, Women, and Voting Citizen - the previous analysis yielded Private Work, Hispanic and Women. The variables worst explained by this model are once again Office and Mean Commute and we now observe Service instead of Transit.

PrivateWork	Women	VotingCitizen	Hispanic	PublicWork	White	IncomePerCap	Walk	Professional	Poverty
0.9285	0.8796	0.8768	0.8646	0.8417	0.8302	0.8139	0.7949	0.7908	0.7051
Drive	Construction	Black	Unemployment	SelfEmployed	Transit	Service	MeanCommute	Office	
0.6349	0.6229	0.5879	0.5716	0.5101	0.4252	0.4081	0.2542	0.2530	

Office	MeanCommute	Service	Transit	SelfEmployed	Unemployment	Black	Construction	Drive	Poverty
0.74701	0.74576	0.59191	0.57477	0.48992	0.42837	0.41209	0.37715	0.36507	0.29490
Professional	Walk	IncomePerCap	White	PublicWork	Hispanic	VotingCitizen	Women	PrivateWork	
0.20921	0.20514	0.18614	0.16982	0.15826	0.13544	0.12322	0.12043	0.07148	

Table 15. Communalities and uniqueness in PFA

Once again, we estimate the factor scores and check that the factors are uncorrelated. Comparing the correlation matrix between the PCFA and the PFA estimate, we can see that they are very close to each other and that the factor structures exchanged given the methods used (factors 3 and 5) are showcased in the plot as well.

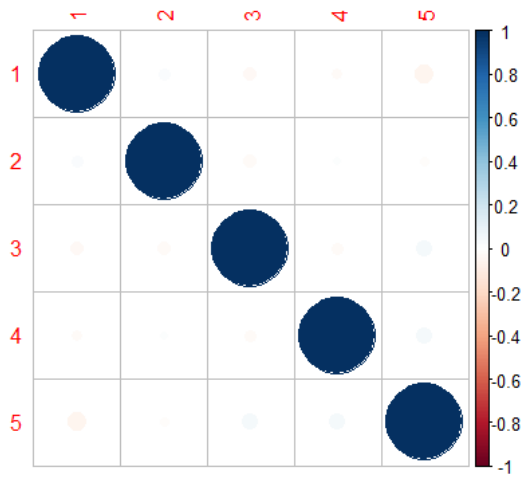


Figure 25. Correlation plot of estimated factor scores.

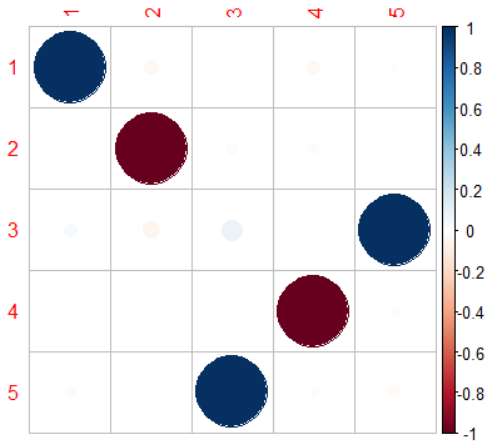


Figure 26. Correlation plot of estimated factor scores of PCFA and PFA.

If we analyze the plot of the correlation among residuals, given the factor scores and the loading matrix , one again we confirm the previously established factors and their associated variables. However, like in the PCFA, it also shows that there are some correlations that the model was not able to explain either.

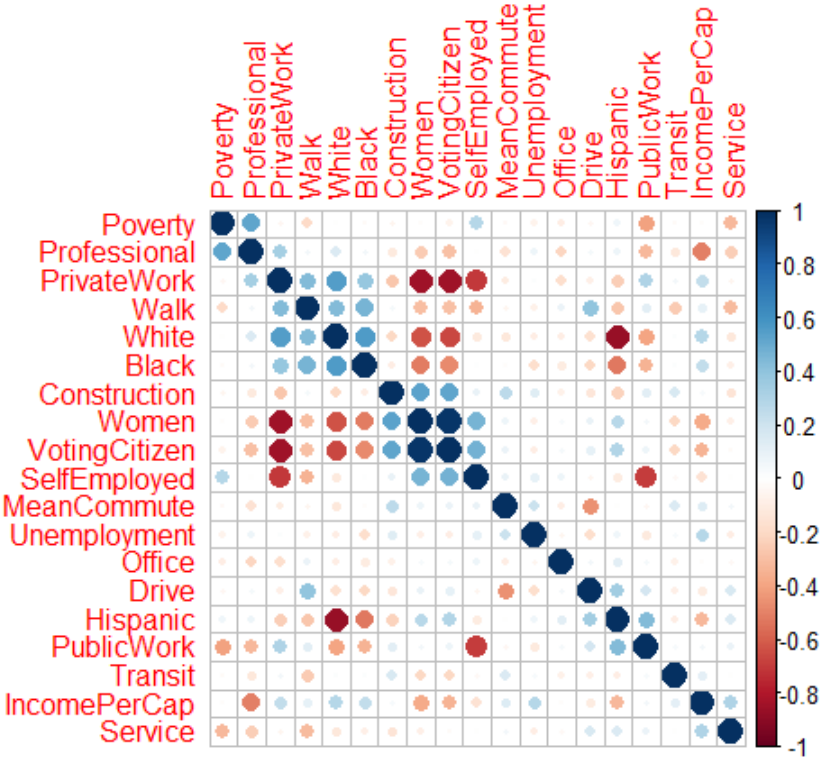


Figure 27. Correlation plot of the estimated residuals given the factor scores and the loading matrix

4.1.3. Maximum Likelihood Estimation

We have previously used non-parametric methods that require no knowledge of the distribution of our data X. In this section we consider the case in which both our variables (x) and the errors (ϵ) are Gaussian distributed, so that we may estimate the MLE factor model.

The estimation of the loading matrix M and the covariance of the errors Σv are obtained through the maximization of the log-likelihood function. Given the data matrix Y with the standardized observations, the likelihood function under the Gaussianity assumption is:

$$L(\varrho_x|Y) = \prod_{i=1}^{n}\left((2\pi)^{-p/2}|\varrho_x|^{-1/2}\exp\left(-\frac{y_{i\cdot}'\varrho_x^{-1}y_{i\cdot}}{2}\right)\right)$$

where $\varrho_x = MM' + \Sigma_\nu$.

Then we can rewrite it as,

$$L(M,\Sigma_\nu|Y) = \prod_{i=1}^{n}\left((2\pi)^{-p/2}\left|MM'+\Sigma_\nu\right|^{-1/2}\exp\left(-\frac{y_{i\cdot}'\left(MM'+\Sigma_\nu\right)^{-1}y_{i\cdot}}{2}\right)\right)$$

Finally, the log-likehood is given by:

$$\ell(M,\Sigma_\nu|Y) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log\left|MM'+\Sigma_\nu\right| - \frac{1}{2}\sum_{i=1}^{n}y_{i\cdot}'\left(MM'+\Sigma_\nu\right)^{-1}y_{i\cdot}$$

Before computing the loading matrix for this estimation, we are going to conduct a likelihood ratio test to consider whether the selected number of factors (r=5) is significant at a 0.05 level. Let us consider the null hypothesis as “the number of factors is r”, and the alternative as “the number of factors is not r”.

H_0 : the number of factors is r
H_1 : the number of factors is not r

The log likelihood ratio is given by:

$$\lambda = n\log\left(\frac{\left|\widehat{M}\widehat{M}'+\widehat{\Sigma}_\nu\right|}{|R_x|}\right) - np + (n-1)\,Tr\left(\left(\widehat{M}\widehat{M}'+\widehat{\Sigma}_\nu\right)^{-1}R_x\right)$$

The p-value obtained for r=5, is zero, which would lead us to reject the null and accept that the number of factors to be considered is not 5. However, testing whether r should take any value from 1 to 9, we always reject the null. Lambda is very sensitive to the gaussianity assumption, which could explain why when testing the number of factors for this analysis, the null hypothesis is always rejected, and consequently, it can be said that the overall data may be far from gaussianity. Yet we must note that for the completion of this section, we will assume it is distributed as such.

From the loading matrix M, we can see that this method has kept the same factor structure as PFA, but the variables have varied quite a bit. The first factor does not include *Transit* nor *Self-employed*. The second factor has now ignored *Service*, while the third and fourth factors have remained the same. Lastly, factor 5 only includes *Self-Employed*, a variable that was actually part of the first factor through PFA; therefore, this factor has changed completely. This model does not include *Black, Service, Office, Drive, Transit, Walk,* nor *Mean Commute.*

	Factor1	Factor2	Factor3	Factor4	Factor5
Women	0.95861	-0.005759	-0.191829	0.18781	0.07538
Hispanic	0.10893	-0.042012	0.001865	0.66728	-0.10573
White	0.12706	0.223956	0.188747	0.93672	0.12512
Black	0.3018	0.321843	-0.00178	0.4754	0.3257
VotingCitizen	0.96443	-0.005572	-0.183029	0.16501	0.07658
IncomePerCap	0.28748	-0.884657	-0.049175	0.05172	-0.05765
Poverty	-0.03743	0.810576	0.129575	0.14505	0.03664
Professional	0.32623	-0.672532	0.271698	0.05973	-0.14326
Service	0.06876	0.365803	0.275198	0.10191	0.03646
Office	0.45158	0.066029	-0.102514	0.04228	0.11746
Construction	-0.57639	0.212331	0.005068	-0.05469	-0.21244
Drive	0.15576	0.307387	-0.256075	-0.13171	0.35048
Transit	0.4946	-0.206973	0.034491	0.17283	0.06199
Walk	-0.22963	-0.278357	0.2621	-0.06187	-0.24202
MeanCommute	0.26928	0.108797	-0.093776	-0.04532	0.13545
Unemployment	0.24937	0.601971	0.127226	0.18308	0.18332
PublicWork	-0.1808	0.193366	0.94579	0.11137	0.06382
PrivateWork	0.40814	-0.044503	-0.822934	-0.03937	0.38437
SelfEmployed	-0.44768	-0.174642	0.037317	-0.10089	-0.83353

Table 16. Estimation of the matrix M by maximizing the log-likelihood function.

Estimating the factor scores (F) and residuals (N),

$$\hat{F} = Y\hat{\Sigma}_{\nu}^{-1}\hat{M}\left(\hat{M}'\hat{\Sigma}_{\nu}^{-1}\hat{M}\right)^{-1}$$

$$\hat{N} = Y - \hat{F}\hat{M}'$$

We may now estimate the covariance matrix of the errors and plot them against the PFA's. It can be observed that there are in fact some differences.

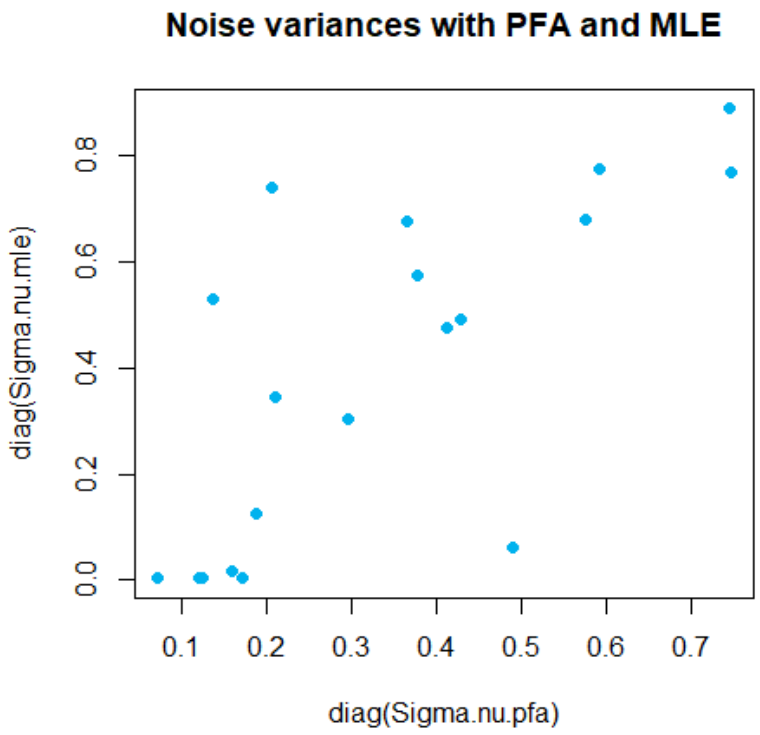


Figure 28. Comparison between PFA and MLE estimates of the covariance matrix of the errors.

Comparing the communalities, the variables that are best explained by the factors are *Voting Citizen, Women* and *Private Work*, the same as with PFA except the order of the first and last variables has been interchanged. The values of the communalities through the MLE are much higher than any of the previous 2 methods. The uniqueness is also much higher under MLE, and although the other has changed compared to PFA, the variables that are worst explained by the factors remained the same: *Mean Commute, Service* and *Office.*

VotingCitizen	Women	PrivateWork	White	PublicWork	SelfEmployed	IncomePerCap	Poverty	Professional	Black
0.9967	0.9967	0.9951	0.9950	0.9811	0.9373	0.8737	0.6976	0.6566	0.5267
Unemployment	Hispanic	Construction	Drive	Transit	Walk	Office	Service	MeanCommute	
0.5079	0.4701	0.4255	0.3245	0.3224	0.2613	0.2344	0.2260	0.1135	

MeanCommute	Service	Office	Walk	Transit	Drive	Construction	Hispanic	Unemployment	Black
0.886457	0.774012	0.765620	0.738688	0.677633	0.675495	0.574539	0.529928	0.492135	0.473250
Professional	Poverty	IncomePerCap	SelfEmployed	PublicWork	White	PrivateWork	Women	VotingCitizen	
0.343364	0.302393	0.126319	0.062745	0.018926	0.004978	0.004934	0.003276	0.003252	

Table 17. Communalities and uniqueness through MLE

Once again, the factors are uncorrelated. Comparing the correlation matrix between PFA and MLE we can see that they are not close, and definitely not as close as PCFA with PFA.

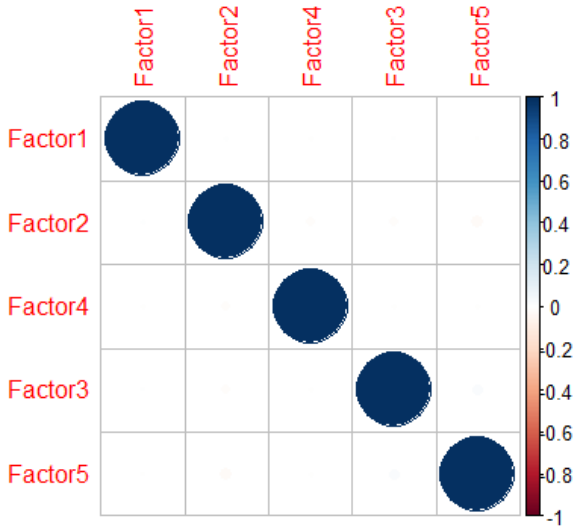


Figure 29. Correlation plot of the estimated factor scores.

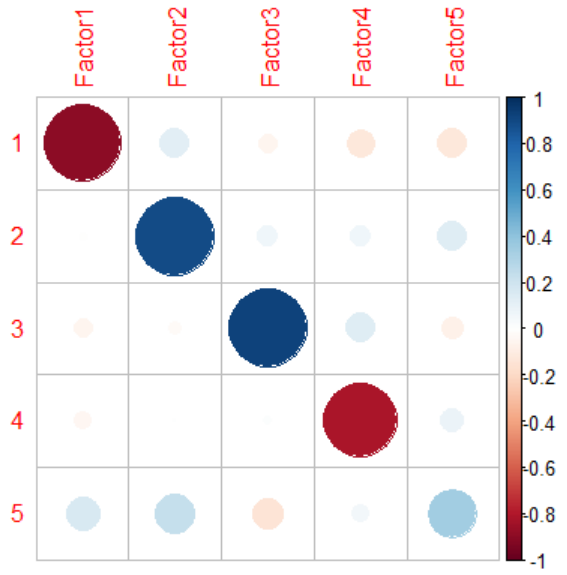


Figure 30. Correlation plot of estimated factor scores of MLE and PFA.

If we analyze the plot of the correlation among residuals, given the factor scores and the loading matrix , once again we visually confirm the previously established factors and their associated variables. However, like in the PFA, it also shows that there are some correlations that the model was not able to explain either.

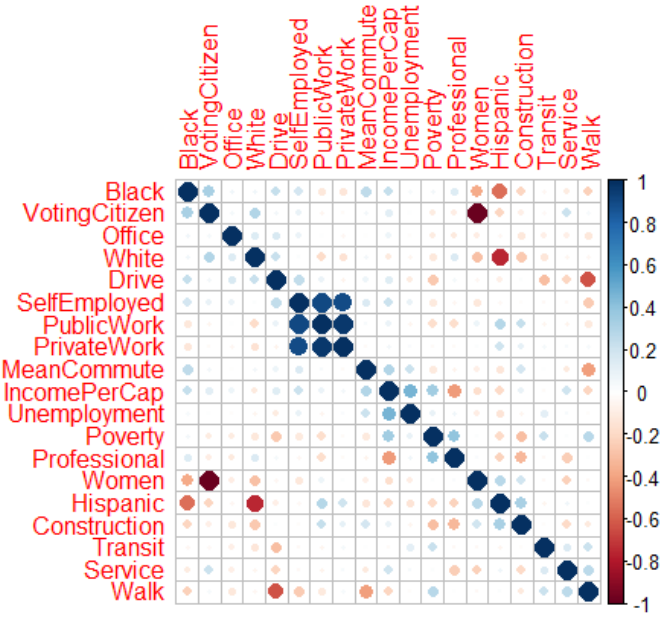


Figure 31. Correlation plot of the estimated residuals given the factor scores and the loading matrix

Although the factor scores, communalities and uniqueness values slightly change from method to method, the variables and latent factors established remained significant, except in the estimations through maximum likelihood - which is not surprising given the violated assumption of Gaussianity. Therefore, the first two non-parametric methods are preferred as we do not need to have any prior knowledge about the respective distribution of the variables, which is advantageous.

5. Cluster Analysis

The most common form of unsupervised learning is Clustering. Clustering algorithms group a set of data points into subsets (clusters). The objective of the algorithm is to create clusters that are coherent internally, but clearly different from each other externally. In other words, elements within a cluster should be as similar as possible, and, simultaneously, as dissimilar as possible with the elements from other clusters.

The most typical clustering methods are: K-mean clustering and Hierarchical clustering. Additionally, there are other as K-medoids or Model-based methods.

5.1. K-mean clustering

It is a clustering algorithm where we attempt to classify our data into *k* number of clusters. The data is mapped into the cluster with its nearest mean of distance.

Before starting, we must have data scaled and none outliers. Also, we must have quantitative features only.

In order to select the most appropriate number of k for our data, we have executed comparisons for few values of k and obtain the one that stabilized value of WSS, that has the highest value of average silhouette or has the maximum Gap statistic.

Therefore, we have implemented it through R Studio with the function *fviz_nbclust.* The result for each method is detailed below in the next figures.

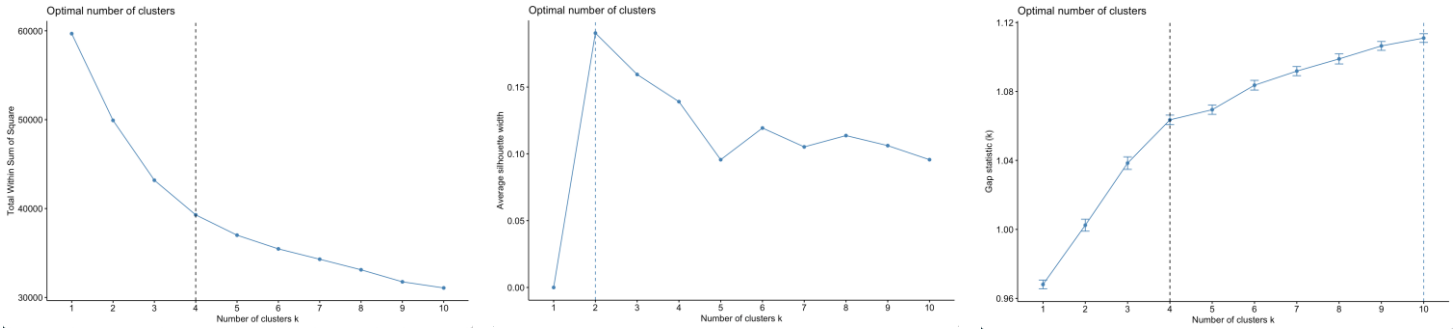


Figure 21. Optimal number of cluster by WSS, Silhouette and Gap statistic methods

The Elbow Curve method (WSS) is helpful because it shows how increasing the number of clusters contribute to separate the clusters in a meaningful way (not in a marginal way). The bend indicates that additional clusters beyond the fourth have little value, so the graph suggests k=4.

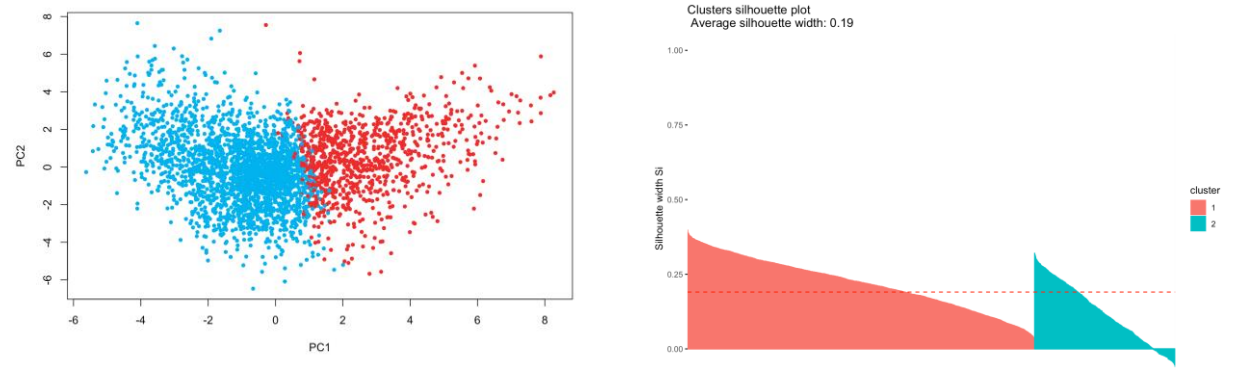
On the contrary, the plot of the Silhouette graph suggest the presence of only 2 clusters since the highest average value of the line take place at k=2. For the next lines, it is important to mention that the higher the value of the average silhouette width and the fewer the number of negative points, the better the method and characteristics executed.

Finally, The Gap statistic plot shows the statistics by number of clusters (k) with standard errors drawn with vertical segments and the optimal value of kmarked with a vertical dashed blue line. According to the plot, k should be 10, but it is because the algorithm didn't converge in 10 iterations, so, let's select it through a visual way. In k=4 the line stops rising for the first time with the same slope, so we can assume that the optimal number of clusters could be k=4.

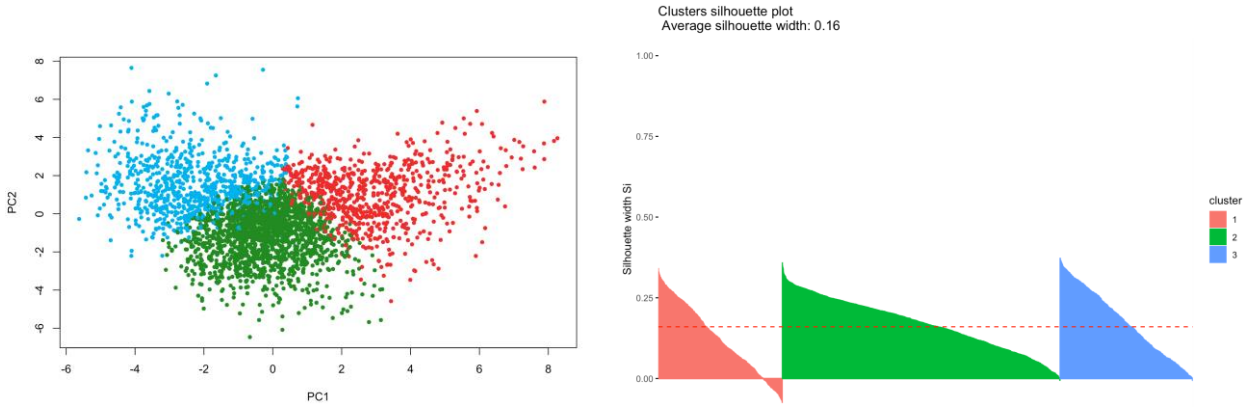
Let's plot the visualization of the clusters for k=2 and k=4. Let's do it for k=3 too, since it is the mean value of the other two.

Then, in order to determine whether k=2, k=3 or k=4 is better, we have computed the plot for the average silhouette width and check which has the highest value and less negative points (points that actually belongs to other group than the assigned one).

K=2



K=3



K=4

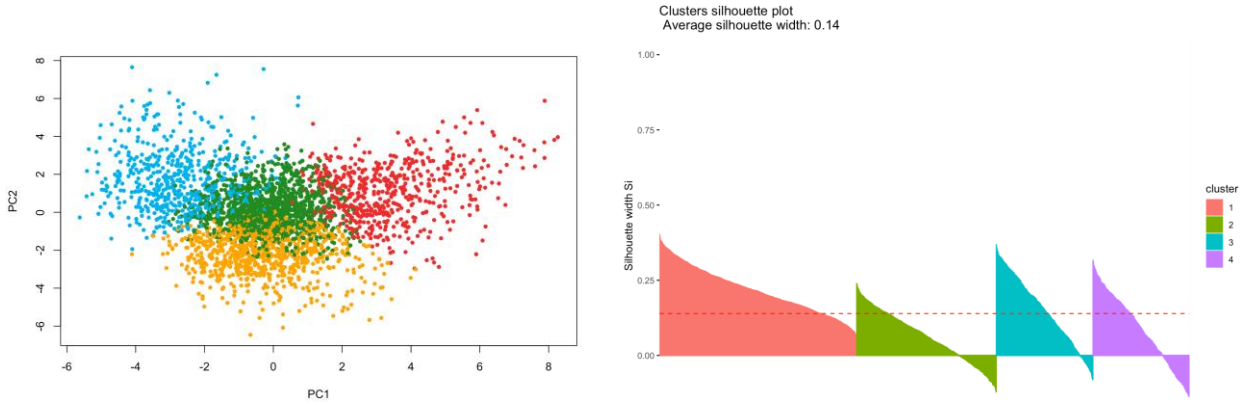


Figure 22. Distribution of the clusters per each number of k and its Silhouette width plot

As we can see in the results, the optimal number of clusters is two, as it has the highest value for the average silhouette width (0.19) and the less negative points in the graph.

To sum up, let's perform the K-means for the first five principal components of section 3 with k=2,, in order to compare how the dimensionality reduction affects in the clustering algorithm.

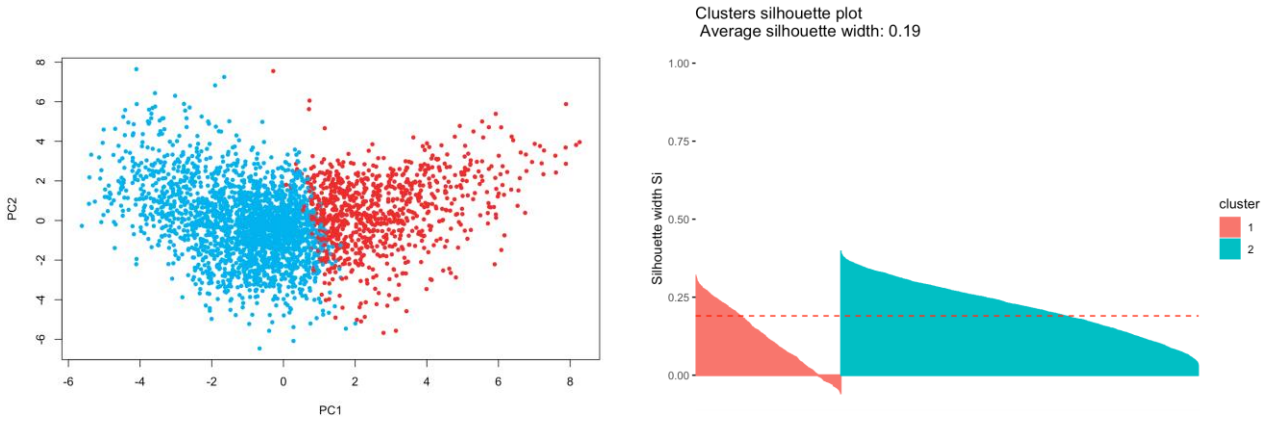


Figure 23. Distribution of the clusters with k=2 for the first five principal components and its Silhouette width plot

As we can see, the results are very similar to the original data (same partition and same value for the average silhouette width (0.19), while PCs only uses 5 variables instead of 19.

They only differ in the silhouette plot where the cluster 1, in red, corresponds to the cluster 2, in blue, in the original set.

Then, according to K-means, k=2 is the best number of clusters for our data set, with a maximum value of 0.19 in terms of ASW.

5.2. Hierarchical clustering

In this case, the clustering is mapped into a hierarchy, reflecting inter-cluster similarities or dissimilarities.

Hierarchy can be bottom up (agglomerative), each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy, or top down (divisive), all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

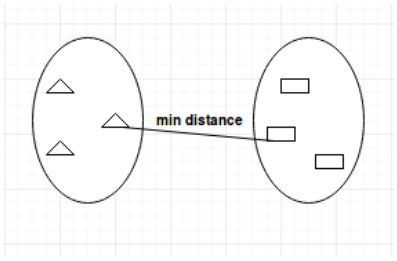
The similarity between the clusters is often calculated from the dissimilarity measures in the distance between two clusters. So the larger the distance between two clusters, the better it is. A main difference of this technique versus the K-means is that it can virtually handle any distance metric while k-means rely on euclidean distances.

To begin with, we have selected the Euclidean distance as a metric for calculating the dissimilarity measure since it is the one used for the Kmeans and that way we can truly compare both results. Later, we will see how the results change when using other type of distance.

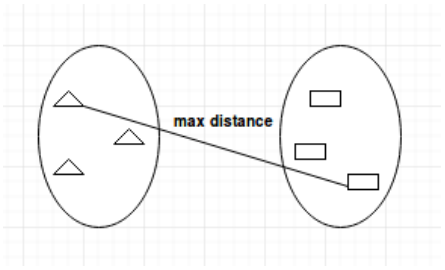
5.2.1. Hierarchical agglomerative clustering

For implementing the agglomerative clustering it has to be selected the linkage method, which are the ones that compute the distance between a new cluster and all other clusters. There are five types of linkage:

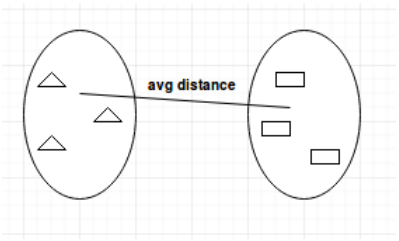
- Single-linkage: calculates the minimum distance between the clusters before merging.
- Complete-linkage: calculates the maximum distance between clusters before merging.
- Average-linkage: calculates the average distance between clusters before merging.
- Centroid-linkage: finds centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between both before merging.



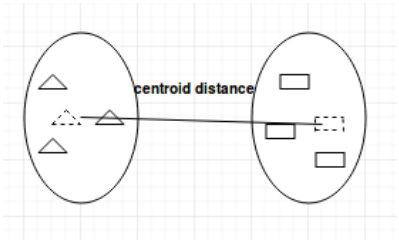
Single-Linkage



Complete-Linkage



Average-Linkage



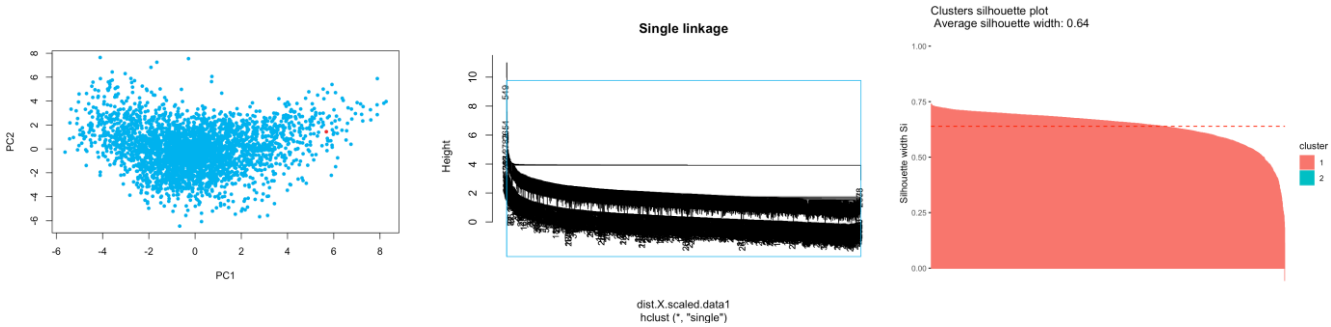
Centroid-Linkage

- Ward-linkage: calculates the Euclidean distance between the sample mean vector of both cluster before merging.

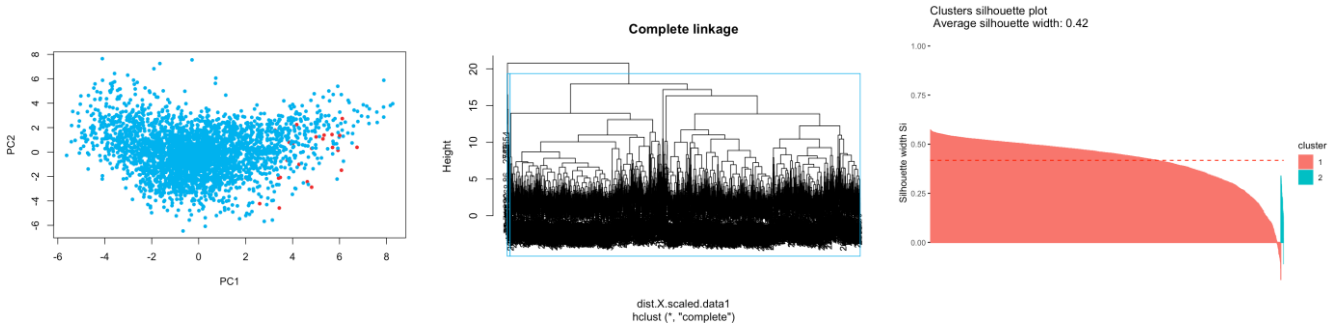
We have computed in R Studio all these methods for two clusters (k=2) in order to compare and choose the best one.

The function used has been *hclust* and we have plotted the dendrograms. The dendrogram shows a tree-like diagram where the distance of split (called height) is shown on the y-axis, and at last, the two clusters are merged into a single cluster where the clustering process stops. Then, we have also plotted the silhouette for obtaining the ASW and see which method performs better.

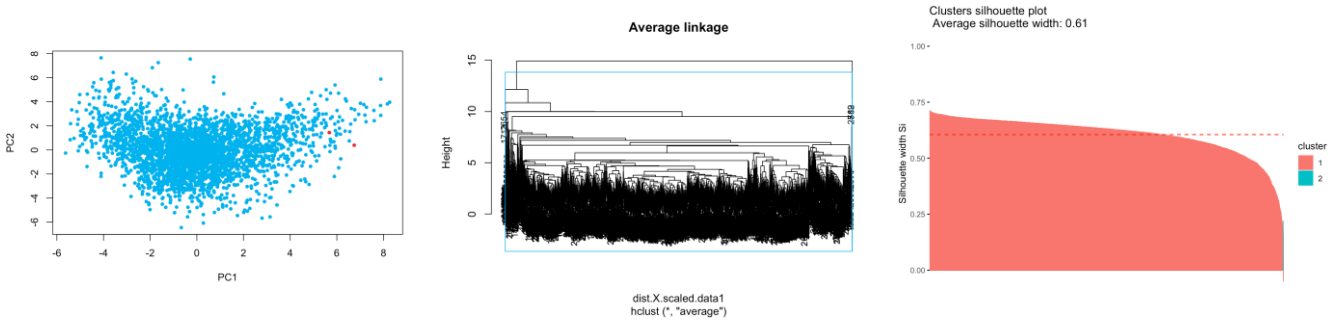
Single-Linkage



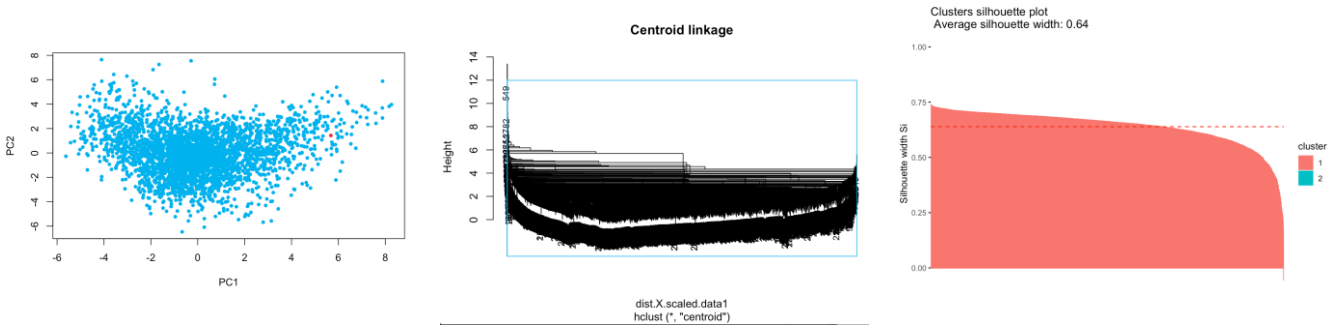
Complete-Linkage



Average-Linkage



Centroid-Linkage



Ward-Linkage

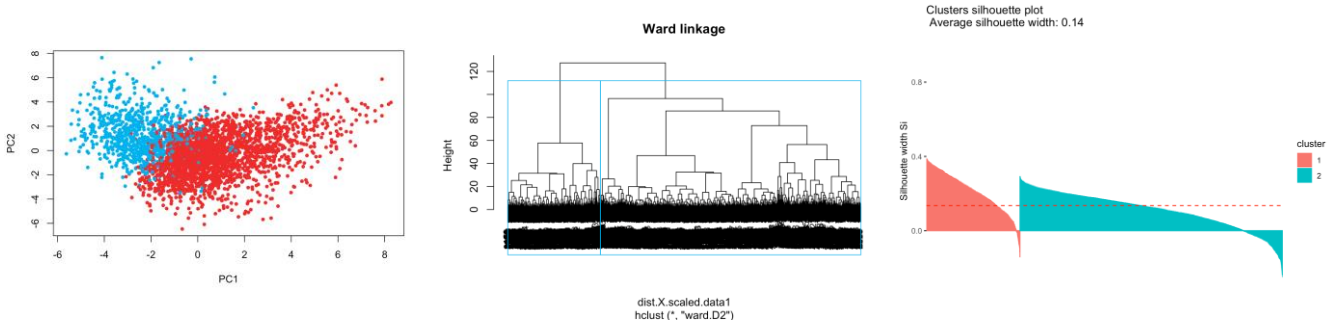
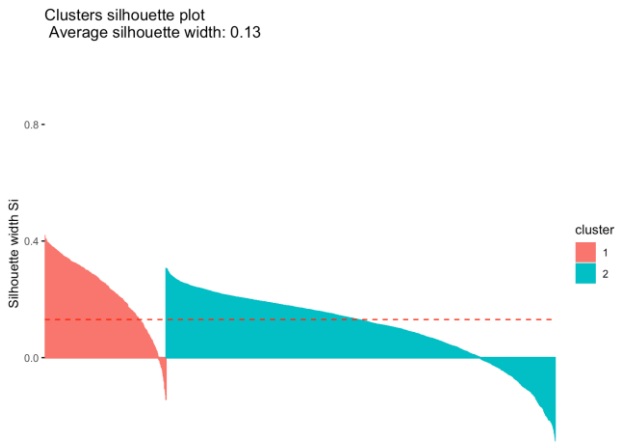


Figure 23. Distribution, dendrograms and ASW plots for k=2 and different linkage methods

As we can see in the plots, the single, average and centroid linkage methods don't work well for our data as they don't separate the data correctly in the two clusters and only select one or two points for the second group. On the contrary, the complete and ward linkage method works better, particularly, the ward method.

In addition, let's obtain the ASW (average silhouette width) for k=2 and for Complete and Ward method with the Manhattan distance as a metric and see how it influences in the result.

Complete-Linkage



Ward-Linkage

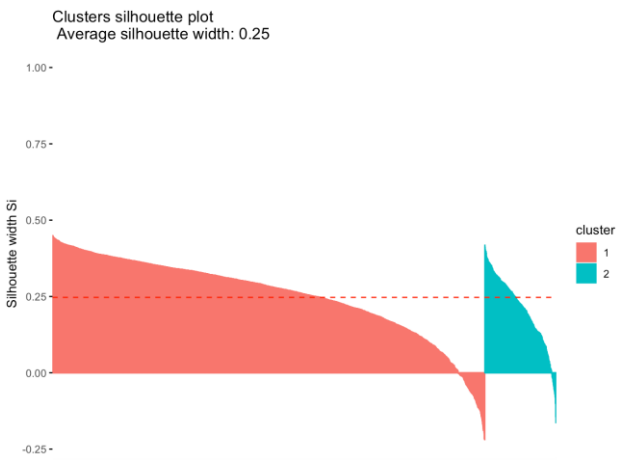


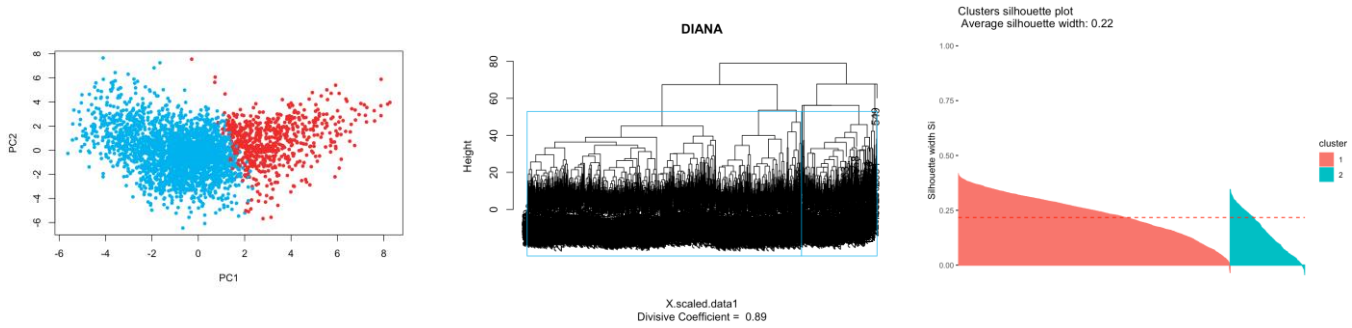
Figure 23. Average Silhouette Width (ASW) plot and values for Manhattan distance

Using the Manhattan distance as a metric we obtain the ASW values of 0.13 for Complete linkage and 0.25 for Ward linkage, which are much higher than the values obtained with the Euclidean distance. Particularly, the Ward method with Manhattan distance works really well as it achieves the highest average silhouette width value in the whole analysis until now (0.25). Therefore, we can assume that this metric is better than the Euclidean distance and we will use it for the next segments of the section. Nevertheless, with these methods there are still many negative values in the division of the clusters.

5.2.2. Hierarchical divisive clustering (DIANA)

For implementing the divisive clustering it has been performance the function *diana* in R Studio with Manhattan distance as a metric.

Then, we have obtained the dendrogram plot where the heights are now the diameters of the clusters before splitting. We have also computed the plot of the clustering distribution for the two groups and the average silhouette width one, in order to determine whether this method is better, or not, than others.



The value of the ASW for the DIANA method is 0.22, although it is not the biggest, is better than in most of the cases. Additionally, in the graph we can see almost none negative points, so, this results looks like the best one of all the performed techniques so far.

5.3. K-medoids clustering (PAM & CLARA)

As the K-means is sensitive to outliers, an alternative method is the k-medoids or partitioning around medoids (PAM) algorithm which is a clustering algorithm reminiscent to the k-means. Both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster, but k-medoids chooses data points as centers (medoids) and can be used with arbitrary distances, while in k-means the center of a cluster is the average between the points in the cluster).

With the function *pamk.best* of R Studio we have calculated the optimal value of k for our data.

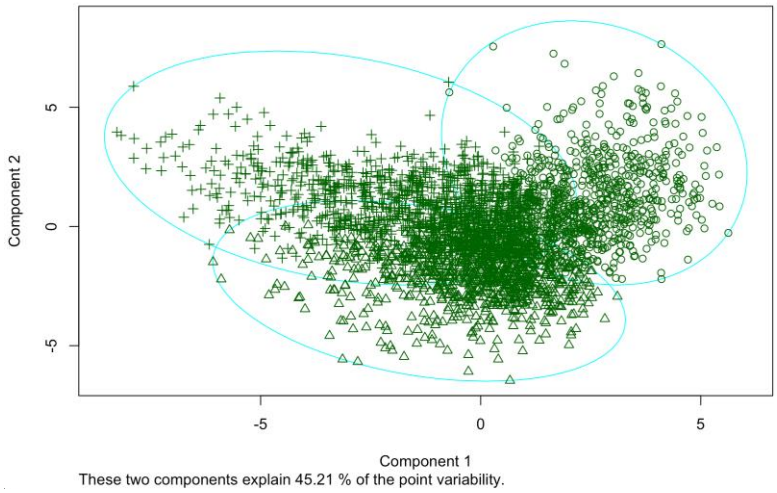
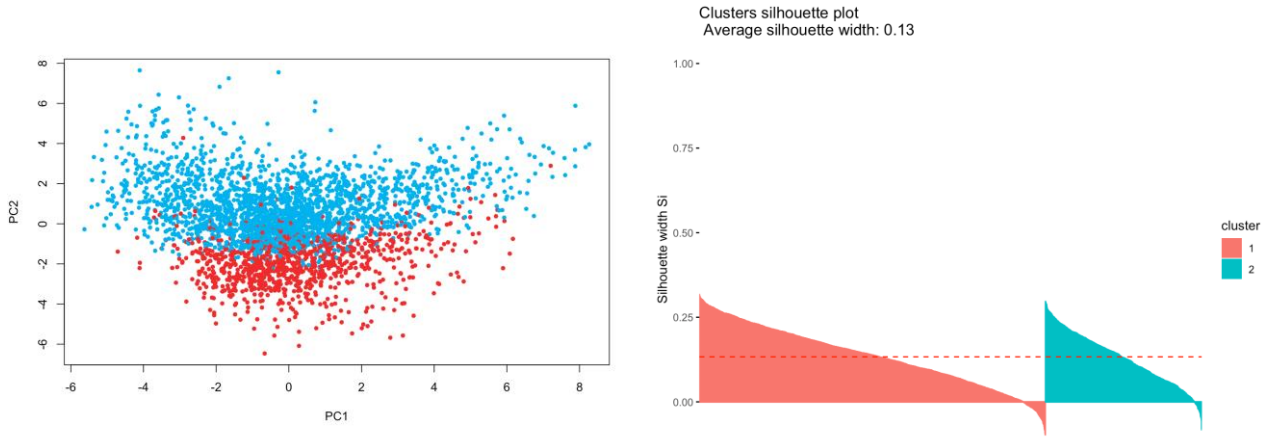


Figure 24. Number of clusters estimated by optimum average silhouette width by PAM algorithm

As we can see in the plot, it suggest a optimal number of 3 clusters.

Let's calculate the average silhouette width with the function *fviz_silhouette* in R Studio (with the Manhattan distance as a metric) for the values of k=2 and k=3, and see which has better performance.

K=2



K=3

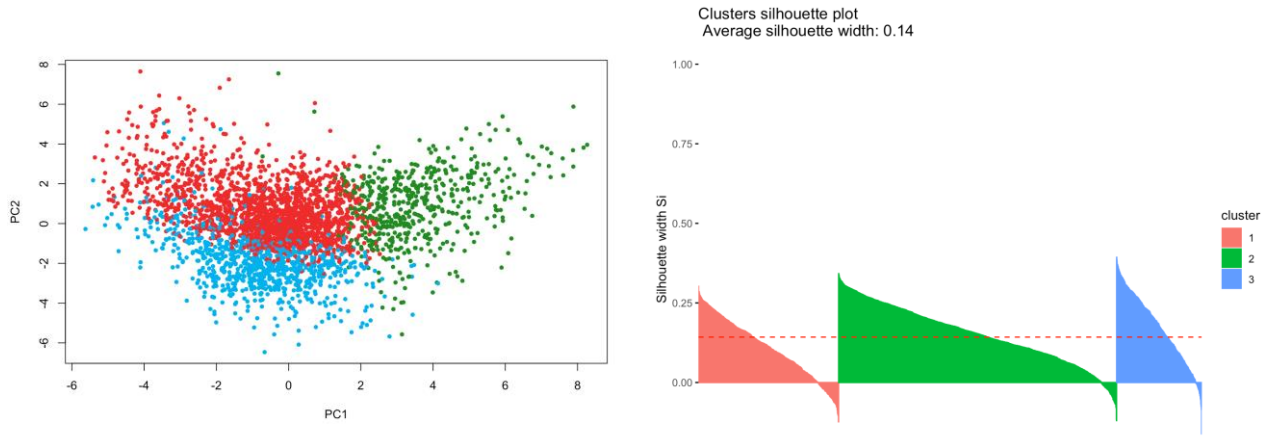


Figure 25. Cluster distributions and average silhouette width values by PAM algorithm

The plots show that, in fact, the value of k=3 has better results for clustering in this case (0.14 for k=3 vs 0.13 for k=2). Nevertheless, it is worse than the one obtained by the K-means algorithm (0.16 for k=3), that could be because the data used for K-means was already out of outliers, so the K-medoids don't make a lot of sense here.

Finally, let's try with the CLARA algorithm which is used for large number of observations (more than 2,000 where we have 3,142) and see which are the results. We have used the same function than before *pamk.best* (with the argument *usepam=FALSE*).

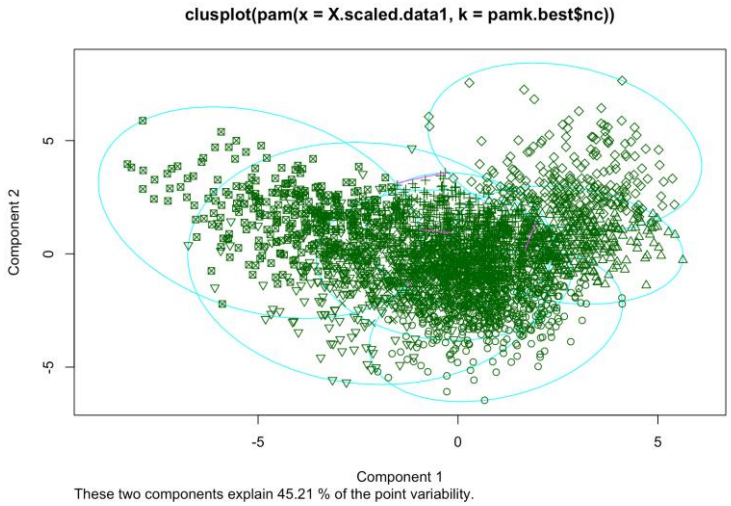


Figure 24. Number of clusters estimated by optimum average silhouette width by CLARA algorithm

CLARA algorithm suggests to have 7 cluster. Let's obtain the comparison between k=2 and k=7 by the ASW values.

K=2

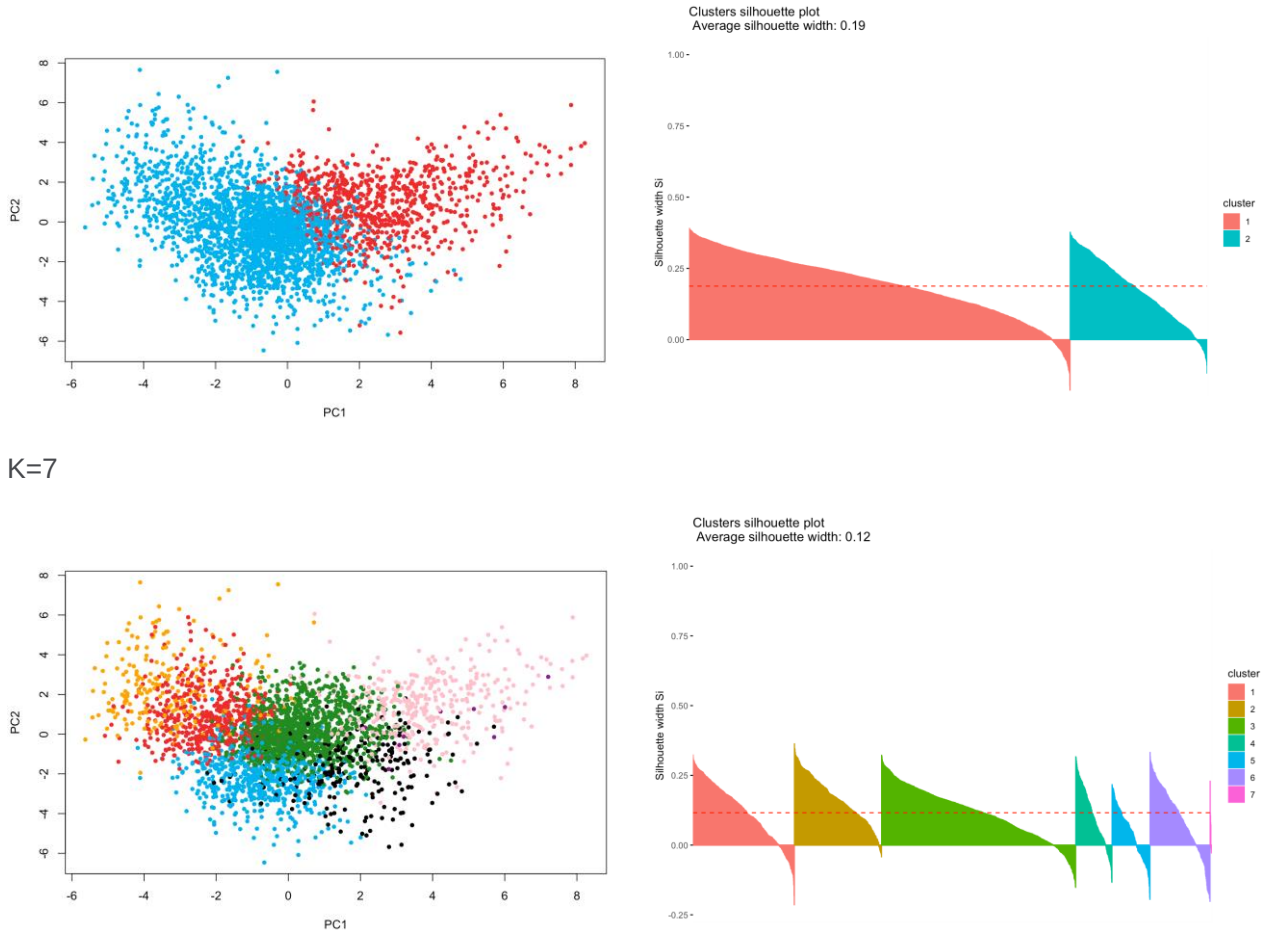


Figure 25. Cluster distributions and average silhouette width values by CLARA algorithm

The result given by the ASW shows that the best number of clusters is still being 2, since the value for k=7 is much worse than the one for k=2, (0.12 for k=7 vs k=0.19 for k=2).

In addition, for k=2 the measure value (ASW) is the same in both techniques (Kmeans and CLARA), but the computational effort is higher in CLARA method, specifically, CLARA is 0.085secs slower.

In conclusion, the result with K-means is still being better than the others partitioning algorithms (PAM and CLARA), with a highest ASW value of 0.19 for k=2.

5.4. Model-Based clustering

Lastly, model-based clustering can also be performed, this method is not a distance-based method, but a probabilistic one.

It usually works well when the observations are generated by different distributions with certain probabilities, then, the observations are assigned to different clusters according to the Bayes Theorem.

The function *mclust* has been used in R Studio in this case, this computes the value of the BIC for all the possible models and we have to select the maximum value of this quantity.

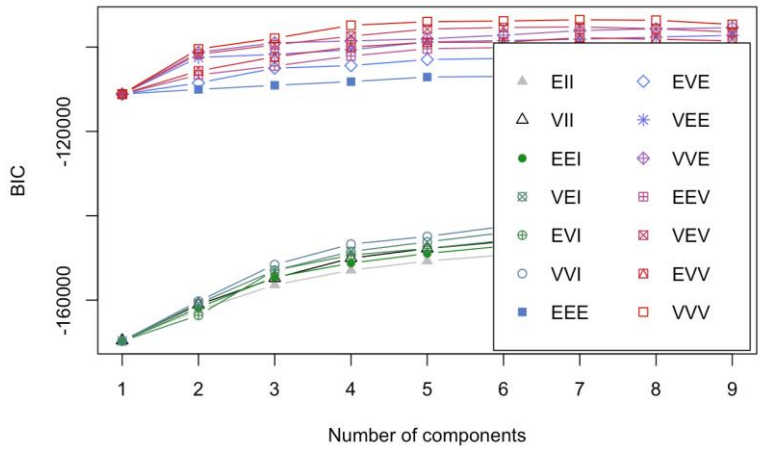


Figure 25. Mclust for the optimal cluster solution

Mclust selects a model with 7 clusters in which the covariance matrices are diagonal and have equal shapes. Then, the mixed probabilities are: 0.20016, 0.14107, 0.14794, 0.12967, 0.12839, 0.08551 and 0.16725.

The distribution of the cluster is detailed in the next plot:

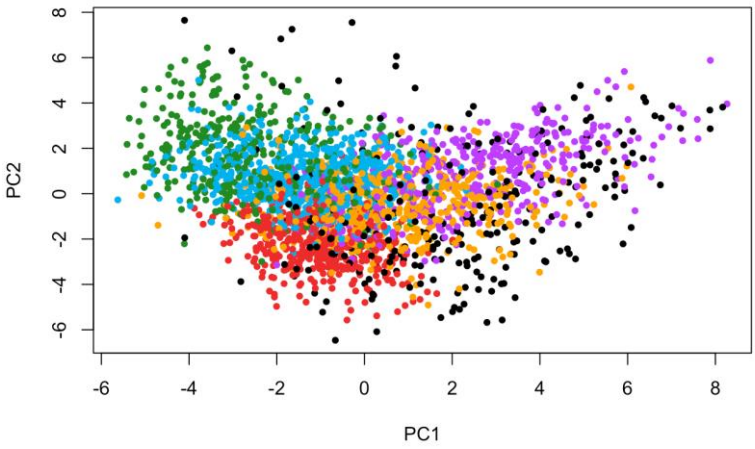


Figure 25. Distribution of the clusters obtained by Model-based method

With only this information we can not assume k=7 is a good number for clustering the data, since the points in the plot are very mixed.

In conclusion, the optimal number of clusters obtained through all the methods has been two (K=2), which gave the highest average silhouette width values (0.19-0.25). On the other hand, we can assume that the best technique for clustering our data is the divisive hierarchical clustering with an average silhouette width value of 0.22 and the lowest number of negative points into the clusters.

The best agglomerative hierarchical method has been the Ward (0.25 of ASW value).

Regarding the distance, the one with the best result has been the Manhattan one, maybe that is why the Hierarchical method has been performed better than the Kmeans, because the Kmeans only compute the euclidean one.

The best partitionally method has been Kmeans, working better than PAM or CLARA.

Finally, mention that there is an option for working with mixed data (quantitative and qualitative variables at the same time) through the Gower distance by the metric *gower,* but we haven't developed this technique, as we haven't qualitative data, except only one "categorical" variable (Counties/States).

6. Conclusions

Because the behaviour of counties differ from one to another, they can be categorized into subgroups in order to obtain a better understanding of the patterns of population, there seem to be some similarities among the regions in which they have been classified. This makes sense, since there are cultural, social and geographical factors that influenced the creation of patterns.

The US, is conformed by 3220 counties, 78 belong to Puerto Rico, which is an independent State in terms of geographical matter, but it is part of the US as a free associated state. Because of historical facts, the rates and information for Puerto Rico diverge from the rest of states in a considerable way. This is the main reason why its counties should not be treated equally with the rest, consequently, these counties were not consider for the analysis.

When faced with a large set of correlated variables, like in this case, principal components allow us to summarize this set with a smaller number of representative variables that collectively explain most of the variability in the original set. Through PCA, it was found a low-dimensional representation of the data set that contains 70% of the variability. The two first PCAs which explained 45.2% of the total variability of the data could be interpreted as “Configuration of the population” and “Wealth”.

The visualization of the first five PCAs dividing counties by the established geographical regions defined by the United States, shows that there are differences between the behavior of the South Region compared to the rest. This pattern can be mostly identified while relating the second and the fourth components, which could be called as “Wealth” and “Hispanics”, then there is evidence to say that these variables could affect the dispersion and common groups within counties which means they have a really high impact on their overall behavior.

While factor analysis may also be helpful in reducing the data into a smaller set of factors it is not is aim, its main goal is to understand the underlying structure of the data and the relationships among the variables given the establishment and identification of latent factors. Even though, the MLE allows for the computation of a wide range of indexes of the goodness of fit of the model and permits statistical significance testing of the factor loadings and correlations, it assumes

that the data follows a multivariate normal distribution which is often violated in practice. Then the non-parametric approaches can be more accurate.

Although the values of the communalities and uniquenesses differ slightly from method to method, all three models agree that the population of women and the percentage of individuals who work for the private sector are best explained by the factors. Other variables on the same note through PCFA and PFA were the population of Hispanics and the number of voting citizens. The significant variables belonged to the "Social determinants", "Ethnicity" and "Business Organization" latent factors. Surprisingly, none of the variables within the latent factor "Economic well-being" and "Socioeconomics" were significant in this form. Some categories of variables, could be omitted completely in order to simplify the analysis of the counties, for example, the type of industry of the jobs, and the type of transportation used to commute as they repeatedly proved to be the worst explained by the factors.

By reviewing both PCA and FA analysis we can jointly conclude that Women are a fundamental piece of information, as well as the number Hispanic minorities. Also, the amount of voting citizens and the rate employed through the private sector. These variables lean towards a more structured socioeconomic component that significantly impacts the behaviour and overall composition of the regions. As an application, these factors may help construct socioeconomic status indices, help us understand and identify battleground States and pinpoint the counties.

The best technique in order to clusterize de data resulted the Hierarchical one, which even though it is computationally expensive, it does not need a definition of a fixed number of clusters in advance. Specifically, the best technique for clustering our data will be the divisive hierarchical clustering which resulted in the lowest number of negative points in the clusters. Moreover, all the clusterization techniques resulted in an optimal of clusters of 2, which can allow us to conclude that this will be the best partition in order to identify groups within counties.

If we compare the results from the clusterization of the two first PCAs divided by the clusterization methods, we are able to see that there are in fact a division of the counties into two major groups. Additionally, this pattern can be contrasted with the division of the PCAs by the region territorial classification, which in fact results not equal with the DIANA method but similar. As we could see in the figures above there is a high proportion of counties from the right bottom that are classified as one group, and this are conformed mostly of counties from the South, while the rest of the counties are what we called the Non-South.

Consequently, the first part of the present project which involved a descriptive analysis resulted in some preliminary conclusions that with the help of the multivariate analysis tools could give us more hits and with higher evidence we can say that there are in fact differences between regions, and that we could find two major groups which will differ significantly in the United States.

7. Bibliography

American Public Transportation Association (2019) .Public Transportation Facts. Retrieved from: https://www.apta.com/news-publications/public-transportation-facts/

Applied Multivariate Statistical Analysis, 5th edition - Wolfgang Karl Härdle, Léopold Simar. Springer Nature Switzerland (2019).

Center for American Progress (2013). The State of Women in America. Retrieved from: https://www.americanprogress.org/issues/women/reports/2013/09/25/74836/the-state-of-women-in-america/

Catalyst (2019). Women in the Workforce – United States: Quick Take. Retrieved from: https://www.catalyst.org/research/women-in-the-workforce-united-states/

U.S. Construction Industry - Statistics & Facts. (2019). Retrieved from: https://www.statista.com/topics/974/construction/.

U.S. CENSUS BUREAU. Census Bureau Regions and Divisions with State FIPS Codes. Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

U.S. CENSUS BUREAU (2017). American Community Survey (2017). Data product COMPARATIVE DEMOGRAPHIC ESTIMATES by county, tables DP03 and DP05. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml

U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU. (2017). *ACS Information guide.* Retrieved the 11th of november of 2019 from: https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf