ETHNICITY AND HOMICIDE IN CALIFORNIA

Mineria de Dades

Dra. Karina Gibert Oliveras

- José Lorente Torres
- José Siqueira Cerqueira
- Carlos de Maqua Rico
- Daniel Ariñez Soriano
- Marc Gallofre Ocaña

11/04/2014

About this study

This study has been realized in the plan of studies of the Data Mining subject of the Facultat d'Informàtica de Barcelona (FIB) which is part of the Universitat Politècnica de Catalunya (UPC). The participants of this study are José Lorente Torres, José Siqueira Cerqueira, Carlos de Maqua Rico, Daniel Ariñez Soriano, Marc Gallofre Ocaña, with the supervision of the subject's professor Dra. Karina Gibert Oliveras. The data of our study were extracted from a repository of online data (open data) where the homicides of the second half of the twentieth century in the California County are shown.

Index:

1. Data Source

This data set examines the relationship between homicide and ethnicity in California during the period 1950 -2000. All the records from the homicides, includes information time, place, location, and cause of the crime for all murder cases in seven California counties. The relationship between victim and accused, and the race, sex, age, and occupation of each are also provided.

2. Data structure and metadata

Prior to the data set study, we will first perform a basic analysis of the sample in our hands. This way it is possible to have a overall idea of what we are looking for, and try to improve or focus our study. Our sample consists of a total of 1316 records with 30 variables, both quantitative and qualitative. These quantitative variables could have NA values (missing), which we will observe carefully.

Our qualitative variables are:

COUNTY	→	Makes reference to the county where the homicide occurred
VICTIM	→	Victim’s name
KILLER	→	Killer’s name
YEAR	→	Year of homicide
MONTH	→	Month of homicide
DAY	→	Day of homicide
HOUR	→	Time frame of the day where occurred the homicide
WEEKDAY	→	Day of the week
VICRACE	→	Victim’s race
VICSEX	→	Victim’s gender
VICOCCUP	→	Victim’s occupation
VICCOND	→	Victim’s condition. Ex: Drunk
ACCURACE	→	Accused’s race
ACCUSEX	→	Accused’s gender
ACCUOCCUP	→	Accused’s occupation
ACCUCOND	→	Accused’s condition
RELATION	→	Relation between victim and accused
CAUSE	→	Killer’s intention
WEAPON	→	Weapon used in the homicide
LOCATION	→	Location where the homicide occurred

Our quantitative variables are:

VICAGE	→	Victim’s age
ACCUAGE	→	Accused’s age
POPULATION	→	County population
NUM.WOMEN	→	Number of women in the county

NUM.MEN	→ Number of men in the county
LAND.AREA.km2.	→ Area in square kilometers of the county
WHITE.ALONE.	→ Percentage of white population in the county
OTHER.RACES.	→ Percentage of non-white population in the county
PERCAPITA.MONEY.INCOME.	→ Per capita money income of the county
HIGH.SCHOOL.GRADUATE.25.	→ Percentage of high school graduates in the county

3. General description of the problem to be analyzed

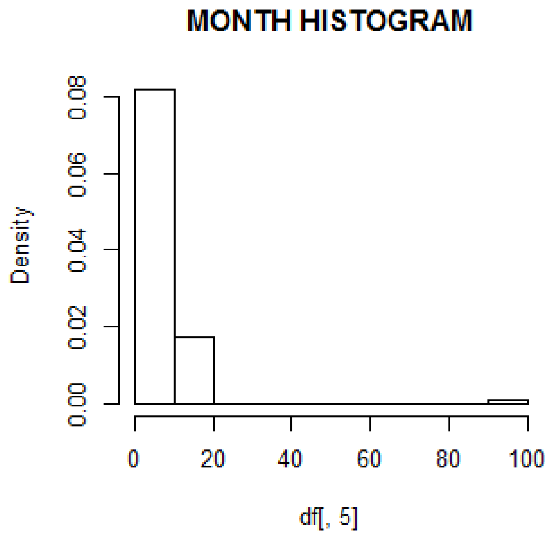
With this study we want to see the connection between the murders in california by an individual, taking into account aspects that refer to the individual as their race, age, gender, and occupation. Similarly will study this same relationship referring to victims of such killings. Also we take into account different aspects relating to the counties where the murders are committed as population, area, and the percentage of graduates. Finally we consider the aspects that reference to the murder itself, as may be the gun, the relationship between victim and accused and the place where it was committed.

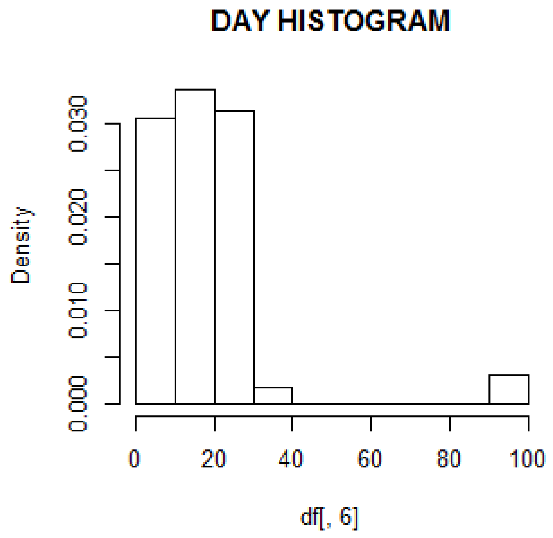
4. Complete Data Mining process proposed

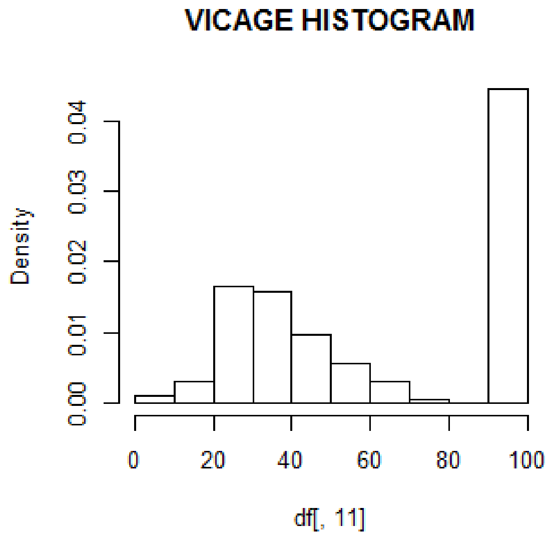
The process that we will be following to realize this study is divided into several processes from data preparation to the generation of final conclusions. In first place we will proceed to clean and prepare the data obtained from our source. Afterwards, we will proceed to perform an analysis of the main components . In sequence, it will be performed an analysis of the multivariable components accompanied by a clustering of the data. Finally we finish with the conclusions extracted from the study of the different analysis performed above.

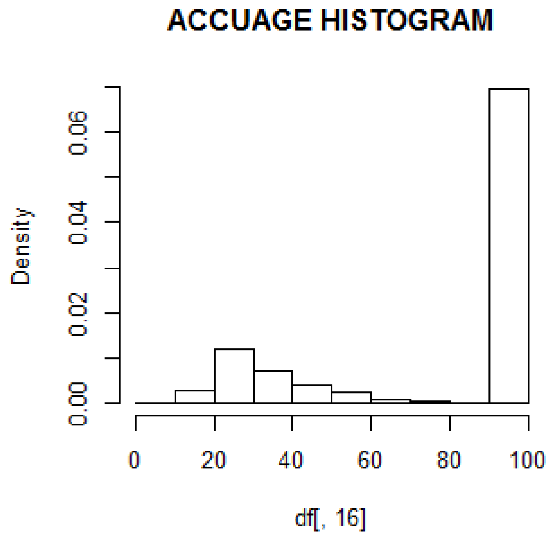
5. Detailed description of Preprocessing and data preparation

In the first part of this phase of preprocessing and preparation of data for our study in the depuration of the initial data frame was done eliminating the missing values in the dictionary that comes together to our data. It is indicated for which variables and what way the missing values are treated giving them a certain value. The variables that contain missing values are: Month, Day, Hour, Vicsex, Vicage, Vicoccup, Viccond, Accusex, Accuage, Accuoccup, Acucond, Relation, Cause, Weapon. The missing values in the qualitative variables were treated as a new class to avoid losing information or distort it. In the case of quantitative variables we proceeded to eliminate these unknown values. In this case we need to clean the quantitative variables Moth, Day, Vicage and Accuage.

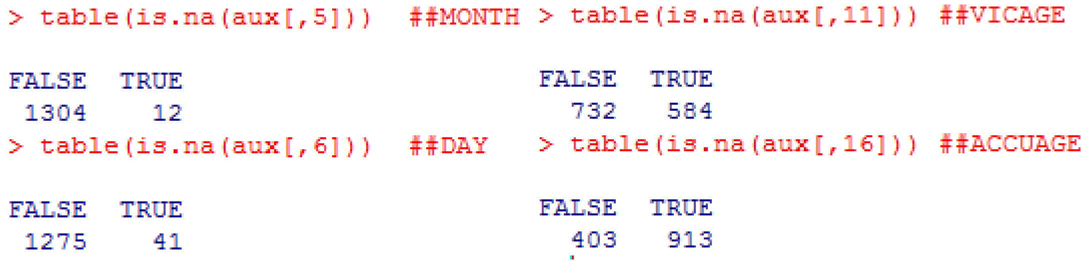








As we can see in the histogram of these four quantitative variables exist a large number of missing values in them, and we must replace them by values that may resemble reality.

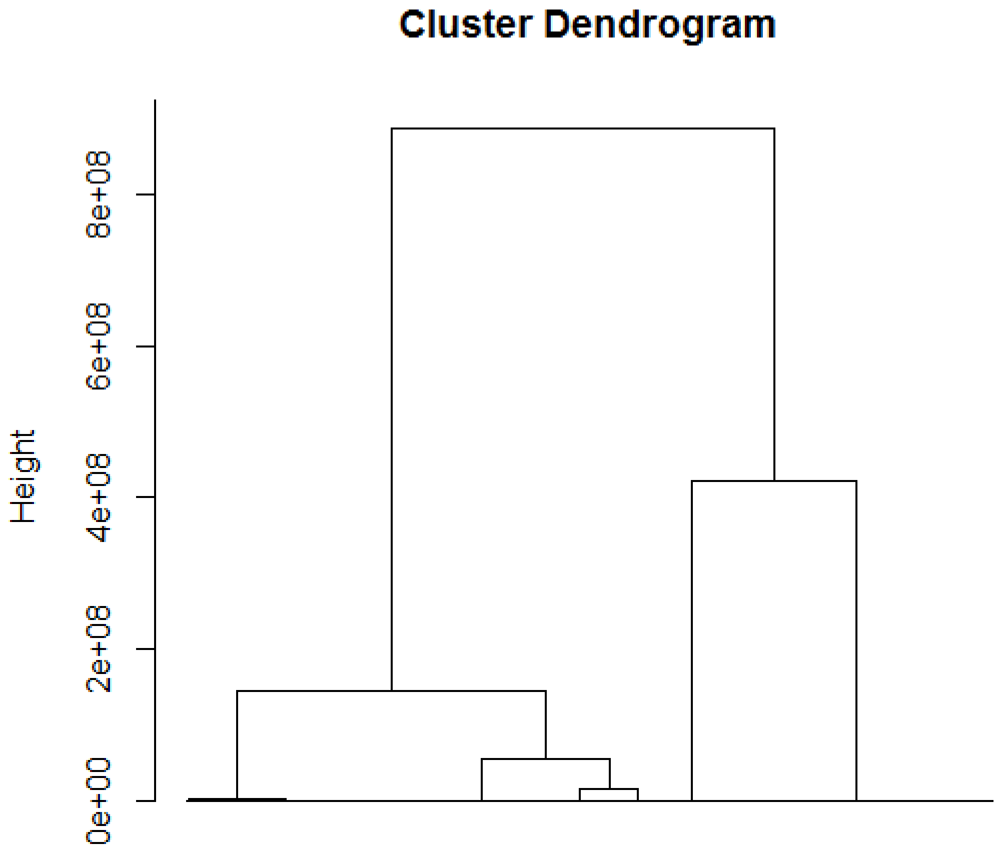


```
> table(is.na(aux[,5]))   ##MONTH > table(is.na(aux[,11])) ##VICAGE

FALSE  TRUE                        FALSE  TRUE
 1304    12                          732   584
> table(is.na(aux[,6]))   ##DAY    > table(is.na(aux[,16])) ##ACCUAGE

FALSE  TRUE                        FALSE  TRUE
 1275    41                          403   913
```

Imputation of NA values process

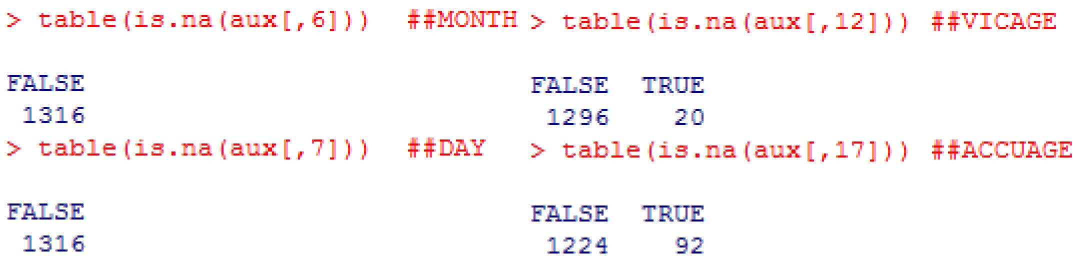
As we can see, our data set contains some missing values, represented by NA. Those are in the numerical variables VICAGE, ACCUAGE, MONTH, YEAR. With the objective to correct and provide a value on these NA, we opted for the intelligent imputation that give us the MIMMI MIMMI [K.Gibert - International Journal of Computer Mathematics (2008), Intelligent Mixed-Multivariate Imputation Missing] method.

To perform the imputation through the MIMMI method, we first seek a few numerical variables without NA or with the minimum value, to perform a hierarchical clustering. These variables are: POPULATION, LAND.AREA.km2., PERCAPITA.MONEY.INCOME., HIGH.SCHOOL.GRADUATE.25. Once selected the variables, we make the hierarchical clustering with the ward method and the euclidean distance. In order to find the values to the imputation of the missing values for groups, a more intelligent way than the variable simple mean.



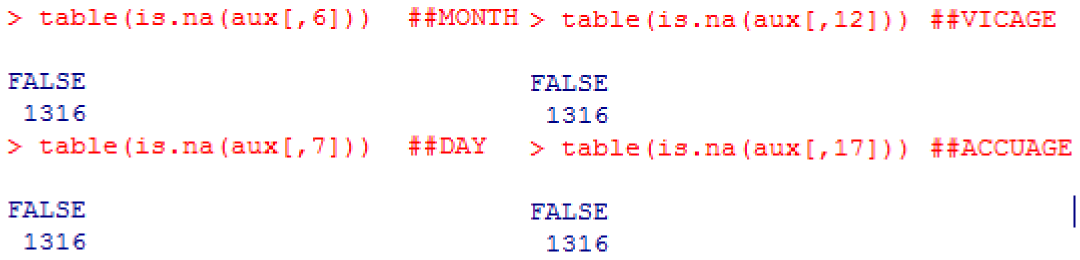
But when we observe the dendogram we realize that separation of groups is done strictly by counties, which we are not interested, as we seek a grouping by more appropriate characteristics of the accused or victim characteristics.

That is why, we opted to do a manual classification based on county and the cause of crime. Thus we keep the classification of the clustering but we add information about the crime, creating more groups. This way we can perform an imputation of the NA values.



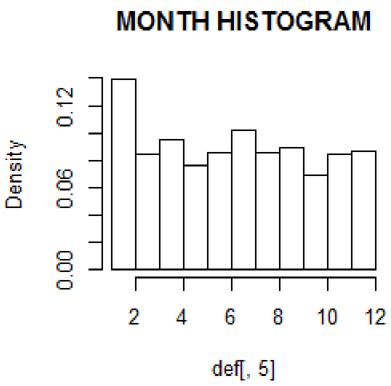
```
> table(is.na(aux[,6]))   ##MONTH > table(is.na(aux[,12])) ##VICAGE

FALSE                              FALSE  TRUE
 1316                               1296    20
> table(is.na(aux[,7]))   ##DAY    > table(is.na(aux[,17])) ##ACCUAGE

FALSE                              FALSE  TRUE
 1316                               1224    92
```

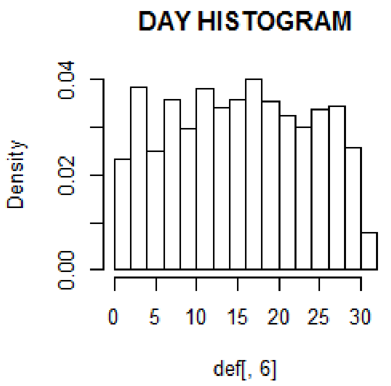
Due to the large amount of groups that we used to make intelligent imputation, we still preserve groups that has only missing values, therefore making it infeasible to impute any value. But being so few values now, we repeat the clustering though the Euclidian distance and Ward method, in Order to impute the last missing values in an intelligent way.

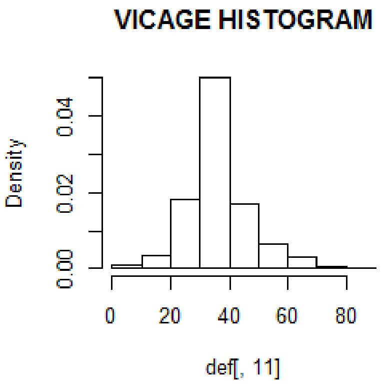


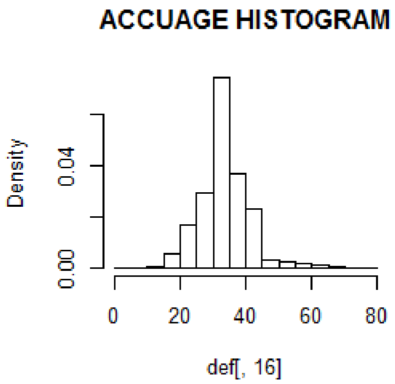
```
> table(is.na(aux[,6]))   ##MONTH > table(is.na(aux[,12])) ##VICAGE

FALSE                              FALSE
 1316                               1316
> table(is.na(aux[,7]))   ##DAY    > table(is.na(aux[,17])) ##ACCUAGE

FALSE                              FALSE
 1316                               1316
```

Now finally, as we can see, there are no longer missing values. And then we can observe the new histograms of these variables. With the variables age with a normal distribution and date with a linear distribution.





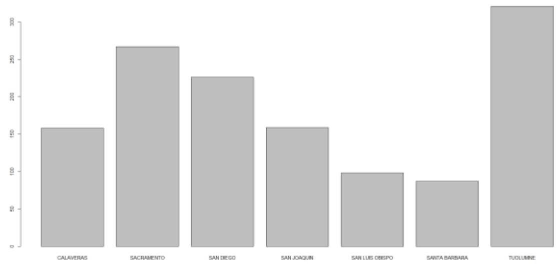




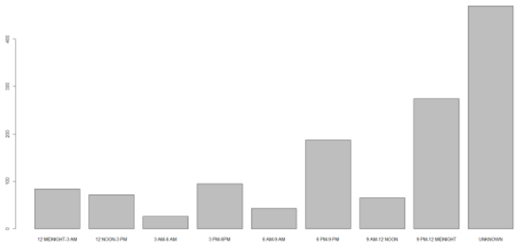
6. Basic descriptive statistics of data sources

Aware of our variables and their types and also considering possible factors that we must keep in mind such as the case of missing values, we will make a first statistical analysis superficially just to have a rough idea of the data we have.

County:

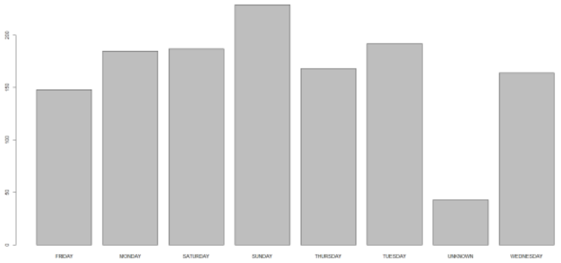


Hour:

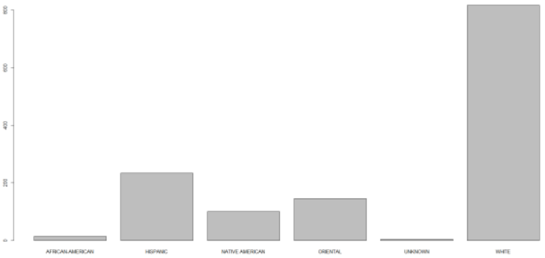


For the variable county we observe that Toulumne beholds the largest amount of homicides, on the other hand the time zone with the largest number of homicides is from 21h to 24h, although the vast majority of them don't know this data.

Weekday:

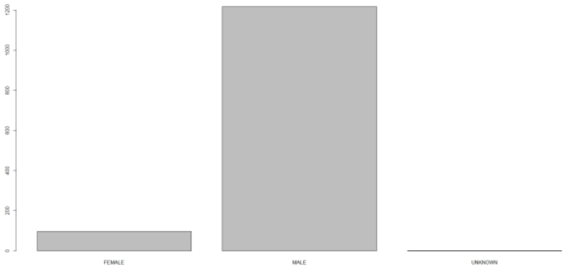


Vicrace:

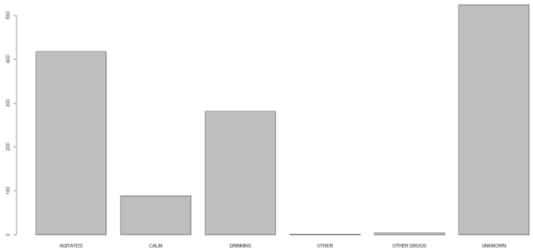


For the variable Weekday we see that the number of homicides is distributed more or less equally by all its values. In the case of variable Vicrace we find that the white race holds most of the murders.

Vicsex:

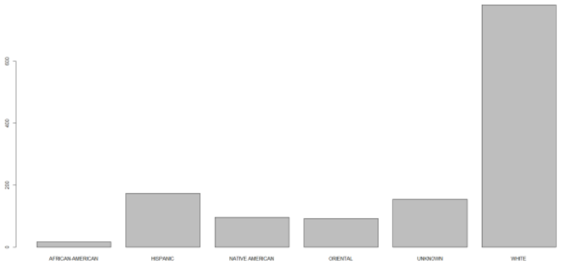


Viccond:

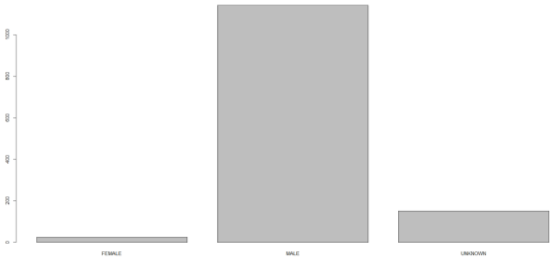


For the variable Vicsex we find that the male sex beholds the largest amount of homicides, on the other hand the variable Viccon with the largest amount of homicides is the agitated state although the vast majority do not know this data.

Accurace:

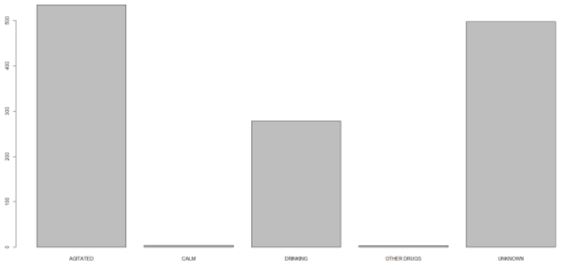


Accusex:

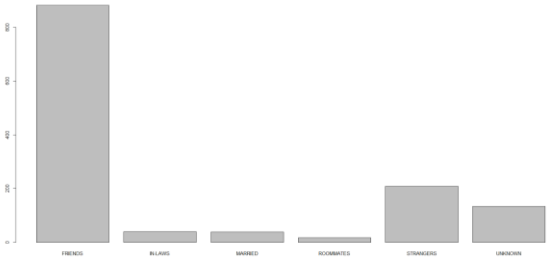


For the variable Accurace we find that the white race beholds the largest amount of homicides, on the other hand the sex with the most number of homicides is male.

Accucond:

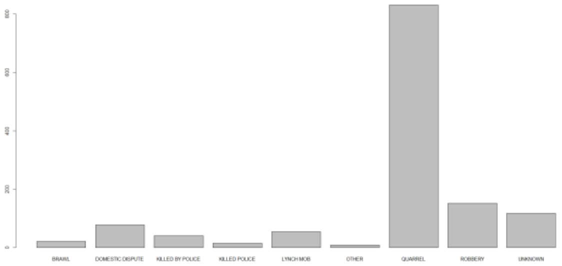


Relation:

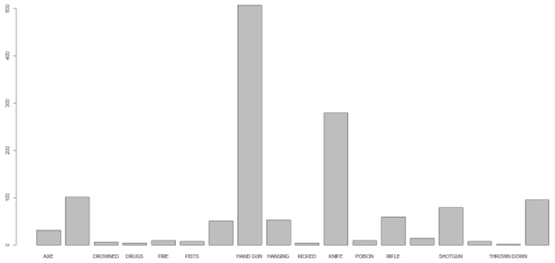


For the variable Accucond we find that Agitated beholds the largest amount of homicides, even though the vast majority of them do not know this data. On the other hand the relation of friends between victim and accused is the one with the largest amount of homicides.

Cause:

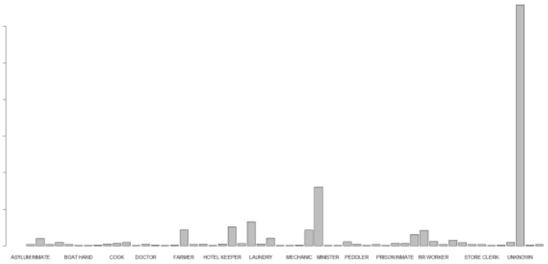


Weapon:

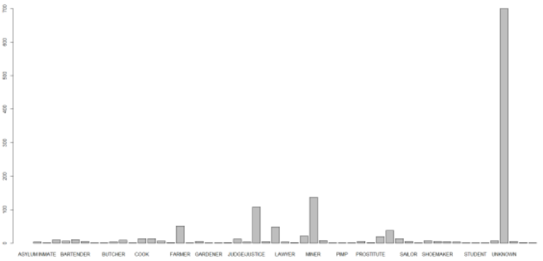


For the variable cause we find that dispute beholds the largest amount of homicides, on the other hand the weapon with largest amount of homicides is .

Vicoccup:

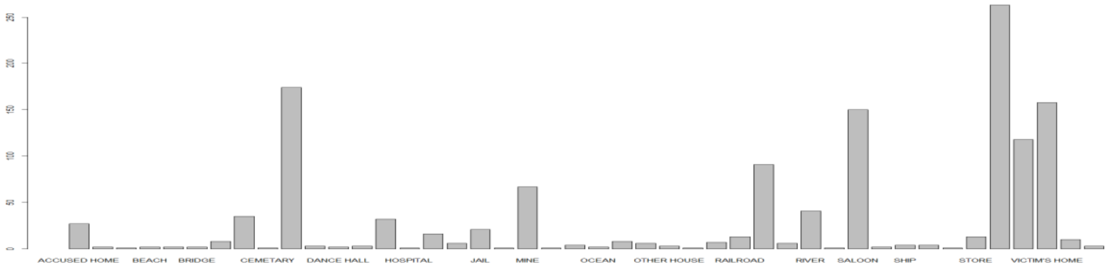


Accuoccu:



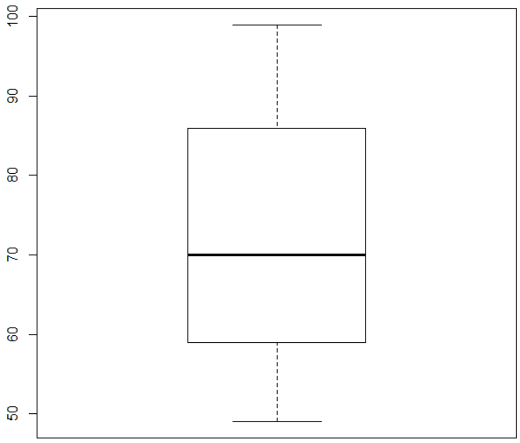
For these two variables that represent occupation both from victim and accused there are too many unknown values.

Location:

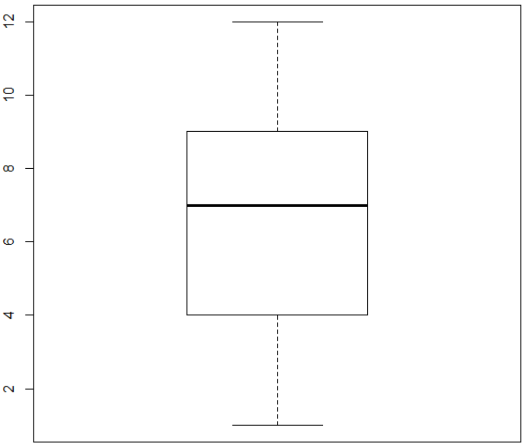


For the variable Location we find that are the streets that holds the largest amount of homicides.

Year:

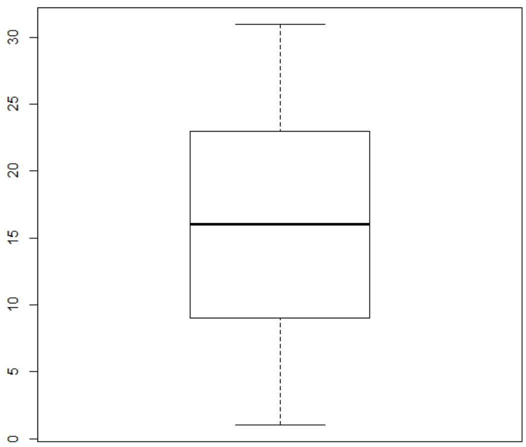


Month:

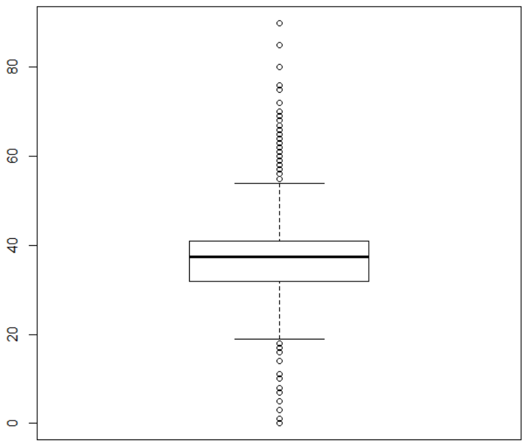


For the variable Year we find that the data are concentrated between the 60 and 85, on the other side the Month variable is concentrated between April and September.

Day:

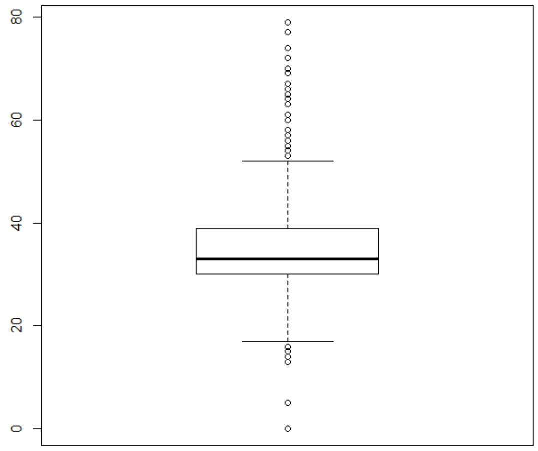


Vicage:

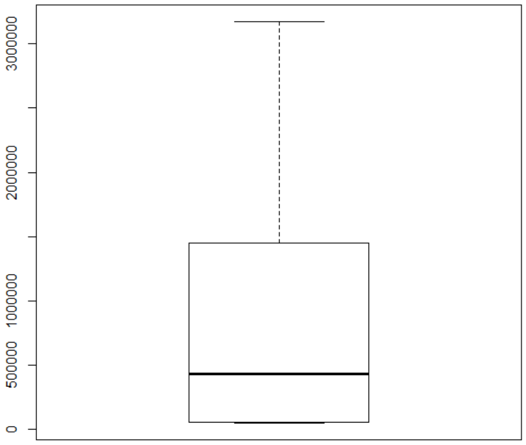


Observing the variable Day we can find that is concentrated on the days 10 and 25 of a month. On the other hand the variable Vicage is concentrated between the values 30 and 40 years old.

Accuage:

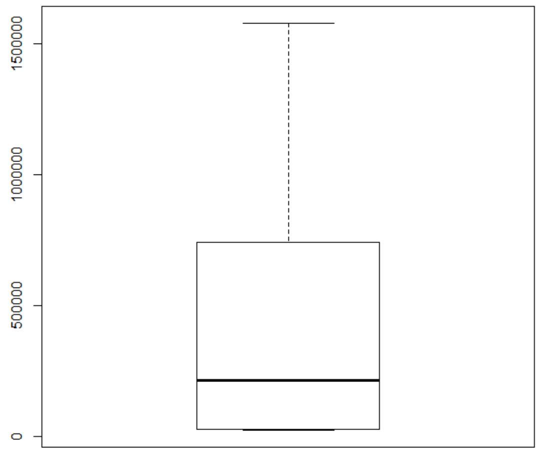


Population:

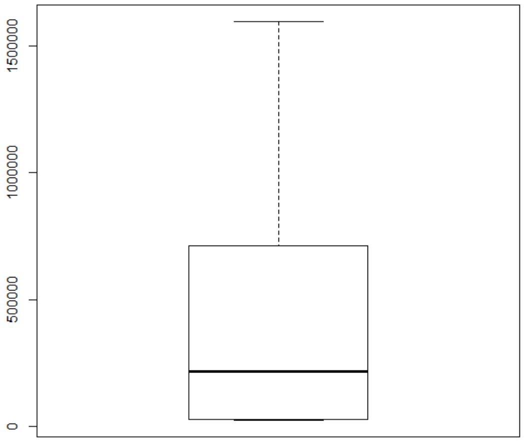


For the variable Accuage we find that also is concentrated between the 30 and 40 years old and for the variable Population we see a concentration between the numbers 0 and one million and four hundred thousand.

Num. Women:

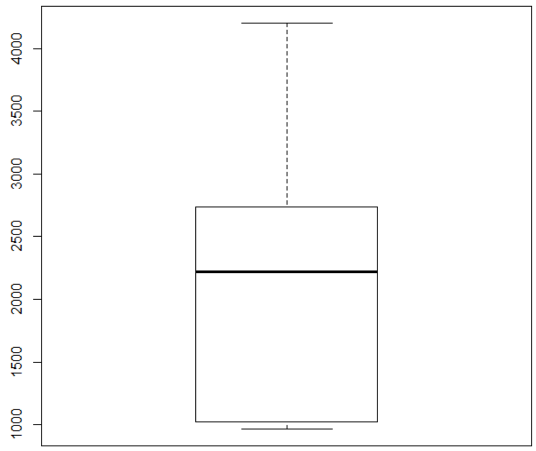


Num. Men

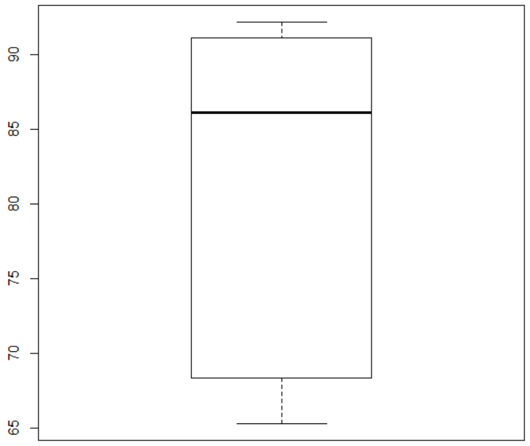


For the variables Num. Women and Num. Men we find that the values are concentrated between 25.000 and 700.00.

Land area:

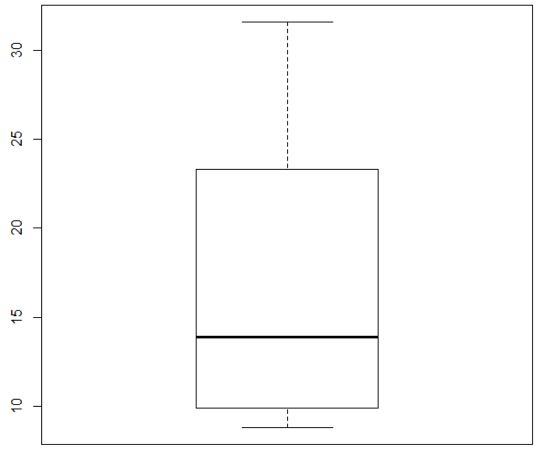


White alone:

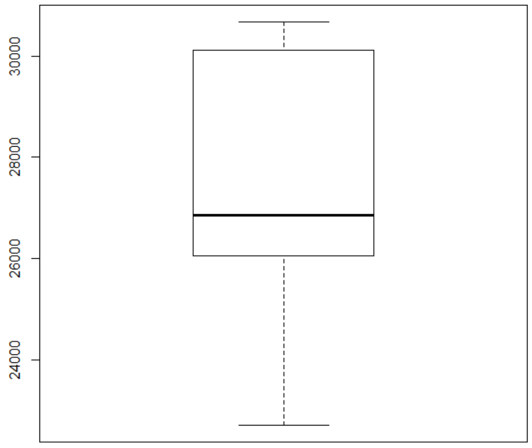


For the variable Land the area of the county the values are concentrated between a thousand and two thousand square kilometers, on the other hand the variable White alone concentrates the values between 70 and 90 percent of the population of the county and a mean of 86% approximately.

Other races:

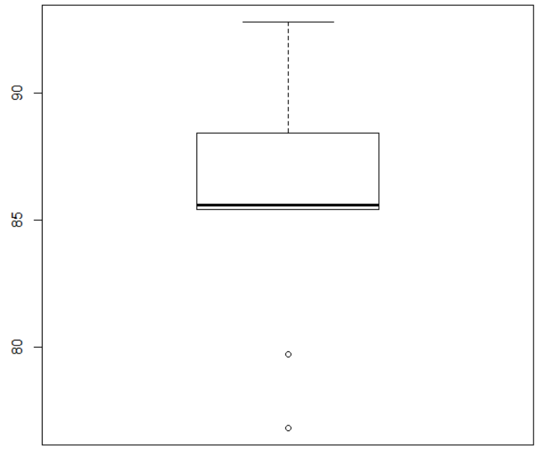


Per capita money income:



For the variable Other races the values are concentrated between 10 and 25 percent with a mean of 16%. On the other hand the values of the variable Percapita money income are concentrated among twenty-six thousand dollars and thirty thousand.

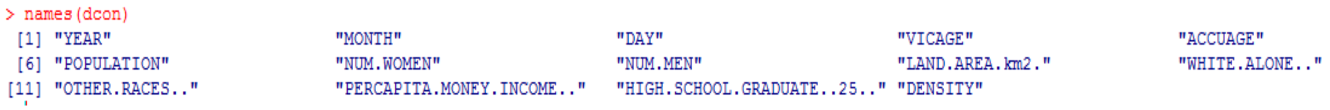
High school graduates:



In the case of the variable High school graduates the values are concentrated between 85% and 90%. Later a more detailed analysis will be shown.

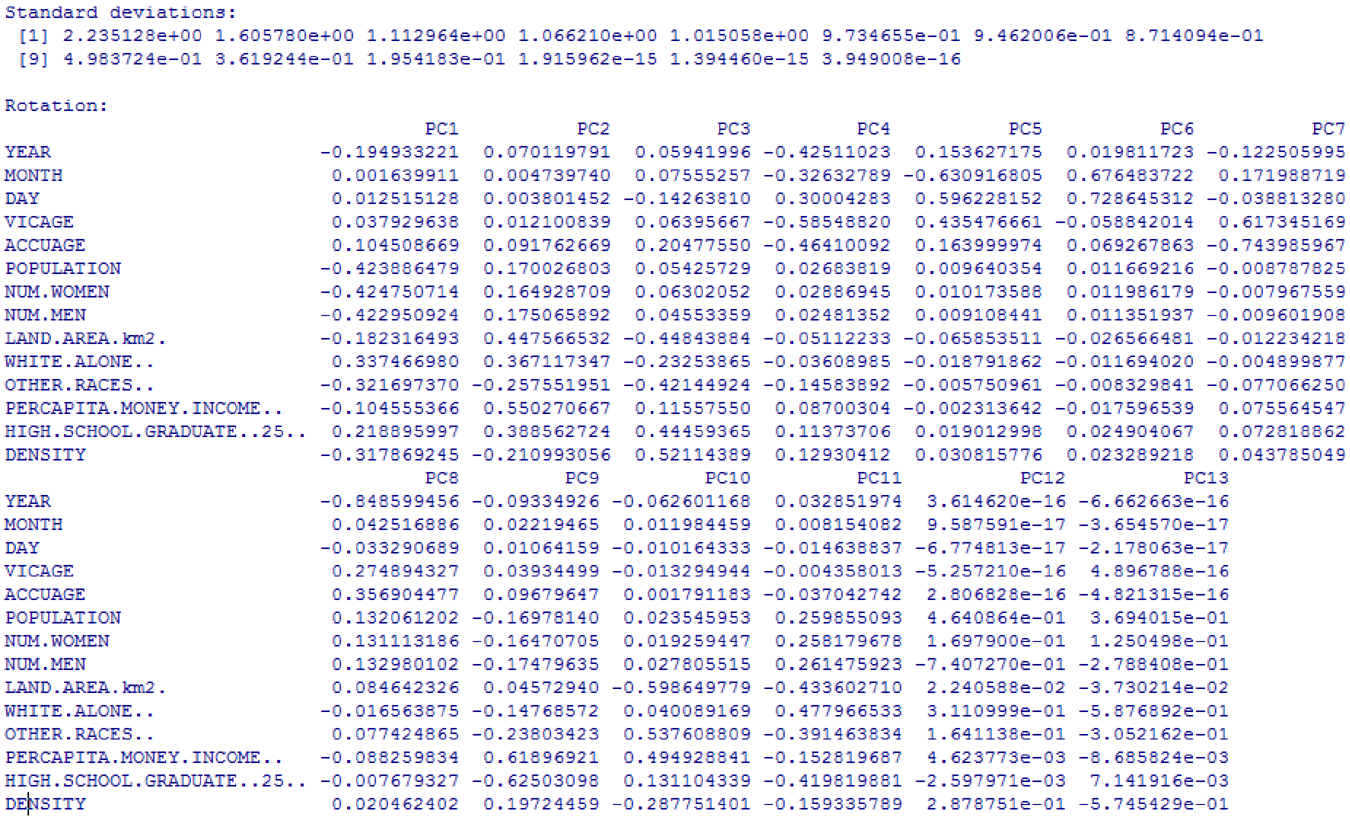
7. ACP analysis

We perform the ACP over the set of our quantitative variables.

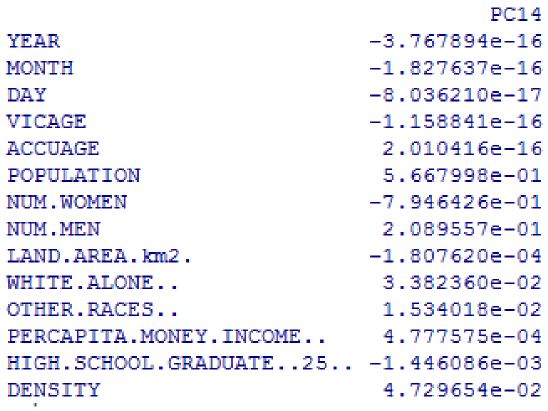


```
> names(dcon)
 [1] "YEAR"                       "MONTH"                      "DAY"                        "VICAGE"                     "ACCUAGE"
 [6] "POPULATION"                 "NUM.WOMEN"                  "NUM.MEN"                    "LAND.AREA.km2."             "WHITE.ALONE.."
[11] "OTHER.RACES.."              "PERCAPITA.MONEY.INCOME.."   "HIGH.SCHOOL.GRADUATE..25.." "DENSITY"
```

The following we study the contribution of the variables to each one of the dimensions.

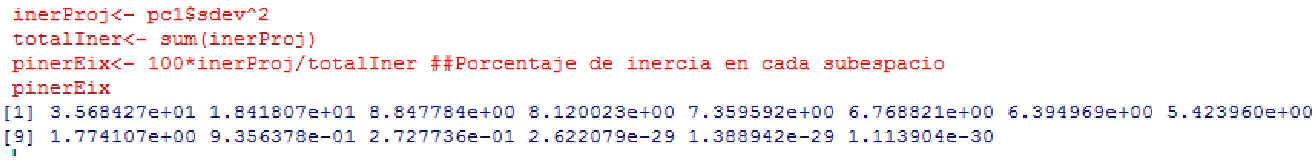


```
Standard deviations:
 [1] 2.235128e+00 1.605780e+00 1.112964e+00 1.066210e+00 1.015058e+00 9.734655e-01 9.462006e-01 8.714094e-01
 [9] 4.983724e-01 3.619244e-01 1.954183e-01 1.915962e-15 1.394460e-15 3.949008e-16

Rotation:
                                   PC1          PC2         PC3         PC4          PC5          PC6          PC7
YEAR                      -0.194933221  0.070119791  0.05941996 -0.42511023  0.153627175  0.019811723 -0.122505995
MONTH                      0.001639911  0.004739740  0.07555257 -0.32632789 -0.630916805  0.676483722  0.171988719
DAY                        0.012515128  0.003801452 -0.14263810  0.30004283  0.596228152  0.728645312 -0.038813280
VICAGE                     0.037929638  0.012100839  0.06395667 -0.58548820  0.435476661 -0.058842014  0.617345169
ACCUAGE                    0.104508669  0.091762669  0.20477550 -0.46410092  0.163999974  0.069267863 -0.743985967
POPULATION                -0.423886479  0.170026803  0.05425729  0.02683819  0.009640354  0.011669216 -0.008787825
NUM.WOMEN                 -0.424750714  0.164928709  0.06302052  0.02886945  0.010173588  0.011986179 -0.007967559
NUM.MEN                   -0.422950924  0.175065892  0.04553359  0.02481352  0.009108441  0.011351937 -0.009601908
LAND.AREA.km2.            -0.182316493  0.447566532 -0.44843884 -0.05112233 -0.065853511 -0.026566481 -0.012234218
WHITE.ALONE..              0.337466980  0.367117347 -0.23253865 -0.03608985 -0.018791862 -0.011694020 -0.004899877
OTHER.RACES..             -0.321697370 -0.257551951 -0.42144924 -0.14583892 -0.005750961 -0.008329841 -0.077066250
PERCAPITA.MONEY.INCOME..  -0.104555366  0.550270667  0.11557550  0.08700304 -0.002313642 -0.017596539  0.075564547
HIGH.SCHOOL.GRADUATE..25..  0.218895997  0.388562724  0.44459365  0.11373706  0.019012998  0.024904067  0.072818862
DENSITY                   -0.317869245 -0.210993056  0.52114389  0.12930412  0.030815776  0.023289218  0.043785049
                                   PC8         PC9         PC10         PC11          PC12          PC13
YEAR                      -0.848599456 -0.09334926 -0.062601168  0.032851974  3.614620e-16 -6.662663e-16
MONTH                      0.042516886  0.02219465  0.011984459  0.008154082  9.587591e-17 -3.654570e-17
DAY                       -0.033290689  0.01064159 -0.010164333 -0.014638837 -6.774813e-17 -2.178063e-17
VICAGE                     0.274894327  0.03934499 -0.013294944 -0.004358013 -5.257210e-16  4.896788e-16
ACCUAGE                    0.356904477  0.09679647  0.001791183 -0.037042742  2.806828e-16 -4.821315e-16
POPULATION                 0.132061202 -0.16978140  0.023545953  0.259855093  4.640864e-01  3.694015e-01
NUM.WOMEN                  0.131113186 -0.16470705  0.019259447  0.258179678  1.697900e-01  1.250498e-01
NUM.MEN                    0.132980102 -0.17479635  0.027805515  0.261475923 -7.407270e-01 -2.788408e-01
LAND.AREA.km2.             0.084642326  0.04572940 -0.598649779 -0.433602710  2.240588e-02 -3.730214e-02
WHITE.ALONE..             -0.016563875 -0.14768572  0.040089169  0.477966533  3.110999e-01 -5.876892e-01
OTHER.RACES..              0.077424865 -0.23803423  0.537608809 -0.391463834  1.641138e-01 -3.052162e-01
PERCAPITA.MONEY.INCOME..  -0.088259834  0.61896921  0.494928841 -0.152819687  4.623773e-03 -8.685824e-03
HIGH.SCHOOL.GRADUATE..25.. -0.007679327 -0.62503098  0.131104339 -0.419819881 -2.597971e-03  7.141916e-03
DENSITY                    0.020462402  0.19724459 -0.287751401 -0.159335789  2.878751e-01 -5.745429e-01
```

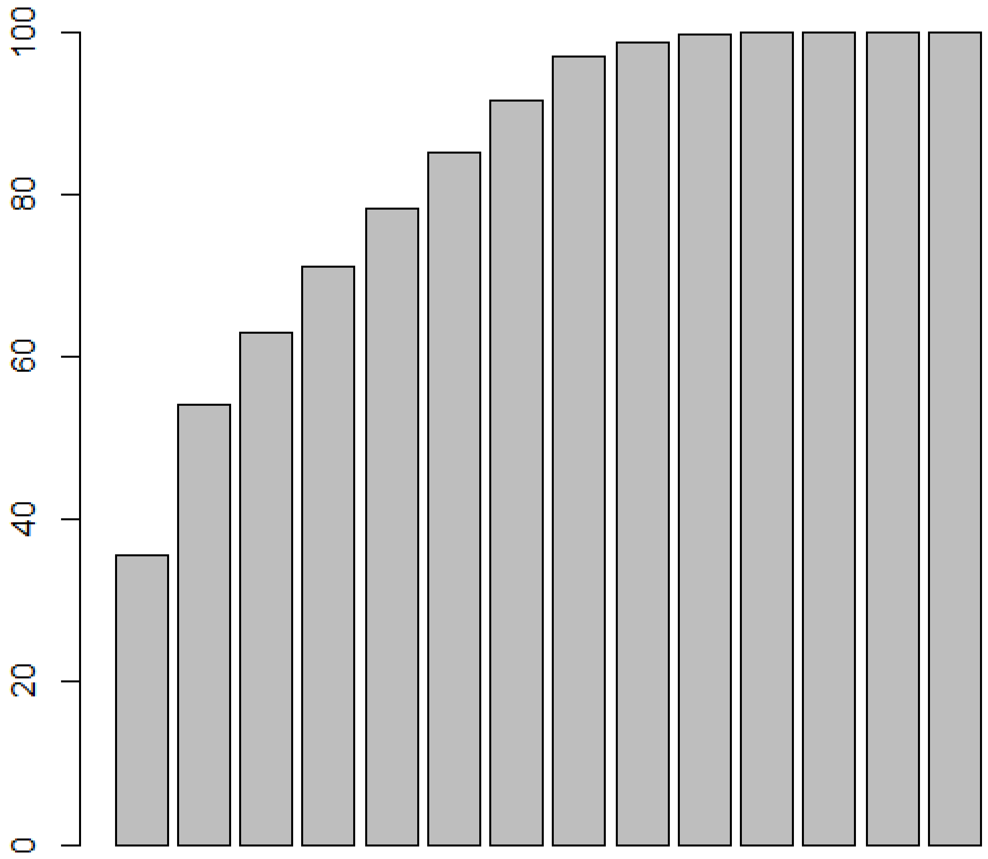


```
                                   PC14
YEAR                      -3.767894e-16
MONTH                     -1.827637e-16
DAY                       -8.036210e-17
VICAGE                    -1.158841e-16
ACCUAGE                    2.010416e-16
POPULATION                 5.667998e-01
NUM.WOMEN                 -7.946426e-01
NUM.MEN                    2.089557e-01
LAND.AREA.km2.            -1.807620e-04
WHITE.ALONE..              3.382360e-02
OTHER.RACES..              1.534018e-02
PERCAPITA.MONEY.INCOME..   4.777575e-04
HIGH.SCHOOL.GRADUATE..25.. -1.446086e-03
DENSITY                    4.729654e-02
```

We study how inertia between different dimensions is distributed.

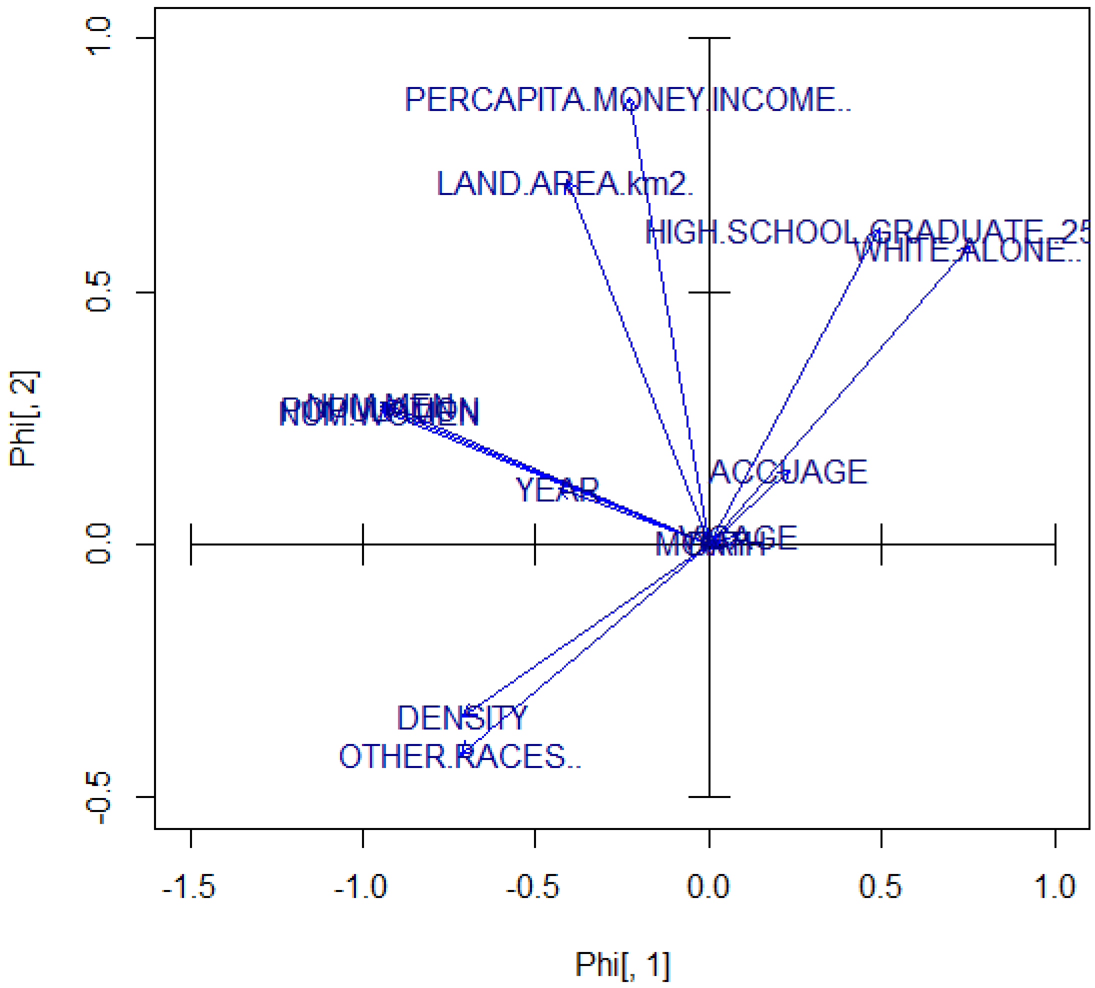


```
 inerProj<- pc1$sdev^2
 totalIner<- sum(inerProj)
 pinerEix<- 100*inerProj/totalIner ##Porcentaje de inercia en cada subespacio
 pinerEix
[1] 3.568427e+01 1.841807e+01 8.847784e+00 8.120023e+00 7.359592e+00 6.768821e+00 6.394969e+00 5.423960e+00
[9] 1.774107e+00 9.356378e-01 2.727736e-01 2.622079e-29 1.388942e-29 1.113904e-30
```



We note that the first five dimensions represent 80% of the total inertia. So we chose these five dimensions as the most representative because they are the ones that provide greater information and in contrast we discard the rest of dimensions by its low significance.

7.1 Variables Factor Map

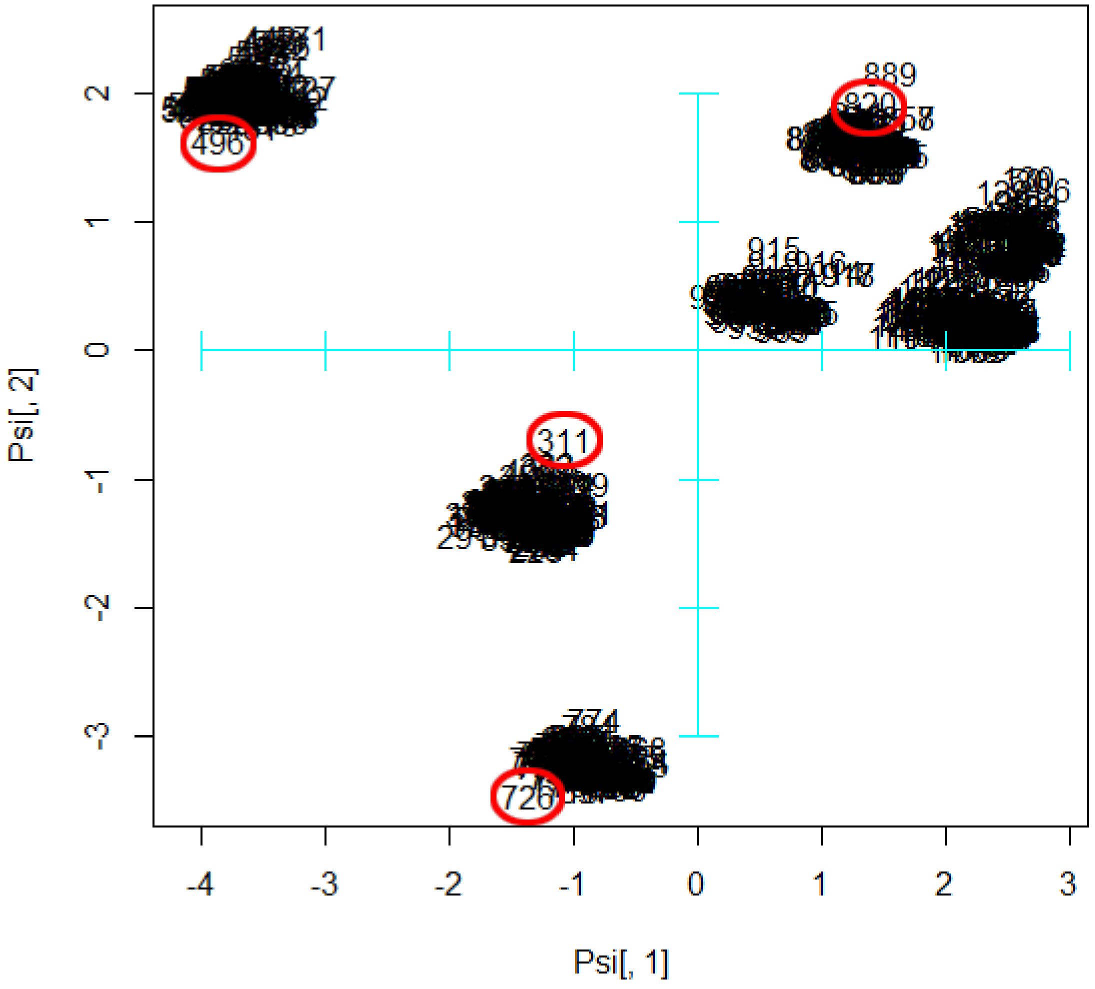


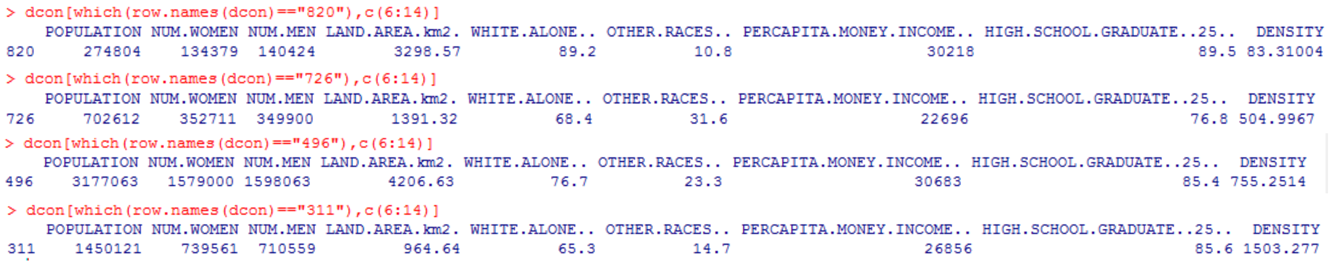
In the Figure we can observe the map of the variables factor and the dimensions 1 and 2. We observe four groups of variables clearly related to each other, those are:

- Density of population - Other races (Population of non white race)
- Higher Education - Caucasian Population
- Surfaces of County - PMI
- Nª Men - Nº Women - Population

It is observed that there are groups of variables that are opposed to each other. Thus we can see that other races (not white) and large population densities are opposed to higher studies and white.

7.2 Coordinates of active elements





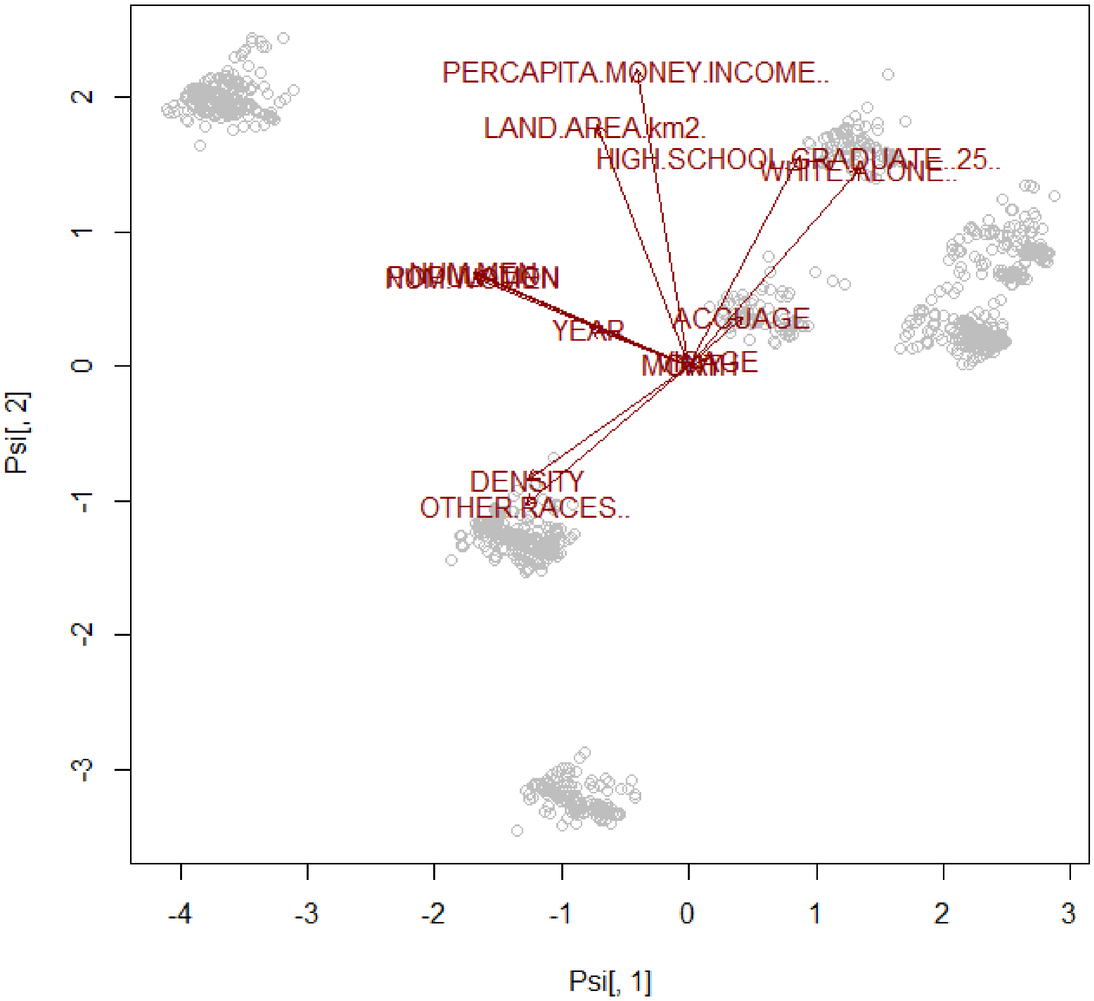
```
> dcon[which(row.names(dcon)=="820"),c(6:14)]
    POPULATION NUM.WOMEN NUM.MEN LAND.AREA.km2. WHITE.ALONE.. OTHER.RACES.. PERCAPITA.MONEY.INCOME.. HIGH.SCHOOL.GRADUATE..25..  DENSITY
820     274804    134379  140424        3298.57          89.2          10.8                    30218                       89.5 83.31004
> dcon[which(row.names(dcon)=="726"),c(6:14)]
    POPULATION NUM.WOMEN NUM.MEN LAND.AREA.km2. WHITE.ALONE.. OTHER.RACES.. PERCAPITA.MONEY.INCOME.. HIGH.SCHOOL.GRADUATE..25..  DENSITY
726     702612    352711  349900        1391.32          68.4          31.6                    22696                       76.8 504.9967
> dcon[which(row.names(dcon)=="496"),c(6:14)]
    POPULATION NUM.WOMEN NUM.MEN LAND.AREA.km2. WHITE.ALONE.. OTHER.RACES.. PERCAPITA.MONEY.INCOME.. HIGH.SCHOOL.GRADUATE..25..  DENSITY
496    3177063   1579000 1598063        4206.63          76.7          23.3                    30683                       85.4 755.2514
> dcon[which(row.names(dcon)=="311"),c(6:14)]
    POPULATION NUM.WOMEN NUM.MEN LAND.AREA.km2. WHITE.ALONE.. OTHER.RACES.. PERCAPITA.MONEY.INCOME.. HIGH.SCHOOL.GRADUATE..25..  DENSITY
311    1450121    739561  710559         964.64          65.3          14.7                    26856                       85.6 1503.277
```

For a good understanding of this graph we have chosen four individuals that we considered representative in order to analyze their values in each variable. So we observe that the individuals 726 and 311 have lower value compared to individual 820 in variables such as higher education or the percentage of white persons. Instead, these individuals (726 and 311) have higher values for the variables population density and other races that the individual 820.

In addition, it is also possible to notice that individuals 820 and 496 have similar values for variables such as higher education White Race making that fall in similar coordinates for the dimension 2, but have major differences in variables such as population, NºMen and Nº Women, which makes the individual 496 to be placed in a position much more to the left in the dimension 1.

7.3 ACP in different dimensions

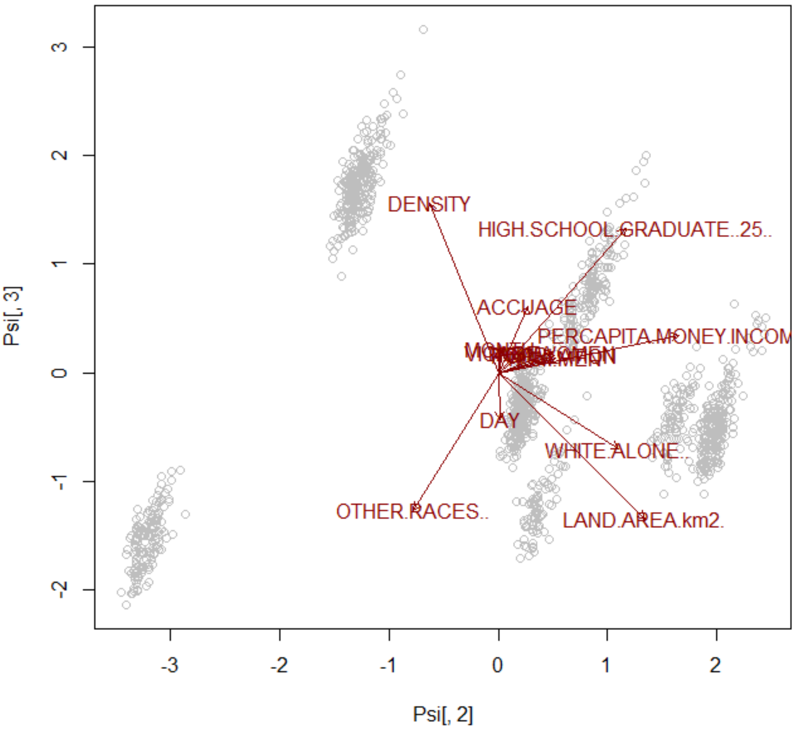
Initially we have chosen the axes 1 and 2 to make our analysis ACP since these are those with greater inertia and thus are more explanatory:



But it has also been contemplated the possibility to use other dimensions, in order to find and discover more information.

7.3.1 Dimensions 2 y 3

Next we conducted a superficial study of map of individuals and variables obtained from the dimensions 2 and 3.



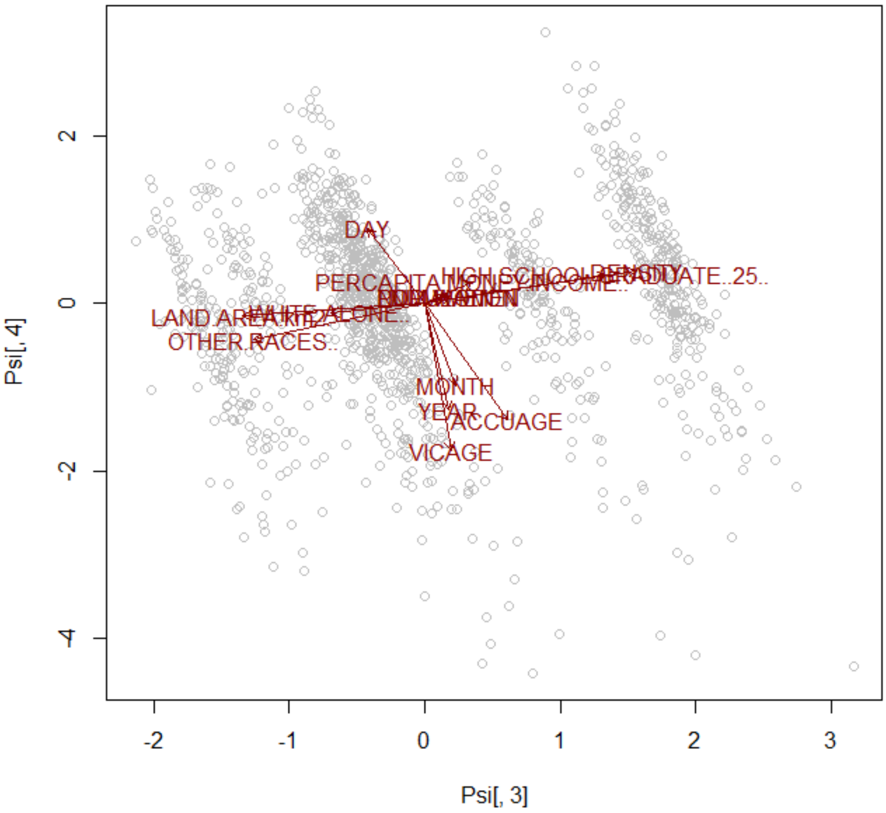
Observed relations between the following groups of variables::

- White alone - Land area
- Density - High.school.graduate - percapita.money.income - accuage

We still observe a contrast between the variables OTHER.RACES and the variable HIGH.SCHOOL.GRADUATE. We also continue to see a clash between the percentage of white race and population density, even that now fades the relationship previously seen between the percentage of different races to the white population and the number of population. Furthermore, we observe that here the variables related to sex and the number of people no longer have a remarkable significance.

7.3.2 Dimensions 3 y 4

Now we realize a superficial study of the map of individuals and variables obtained from the dimensions 3 and 4. continuación realizamos un estudio superficial del mapa de individuos y variables obtenido a partir de las dimensiones 3 y 4.



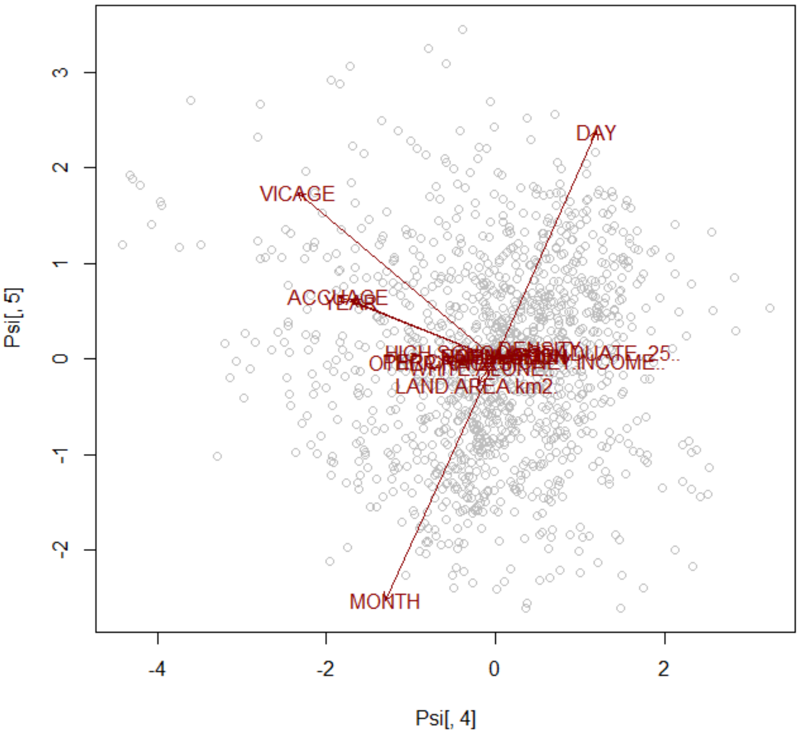
Where we see a great ratio of variables with the third axis, and the following groups of related variables:

- Land.area - other.races - white alone
- Hight school graduet - denstiy
- Moth- year-accuage -vicage

Regarding the axes chosen for the study (1,2), the contrast between the percentage of different races to white and the percentage of white race has faded and now both have been counterposed to the density and level of education of the population. Meanwhile, now the variables date of the crime and the age of the victim and the accused have a remarkable significance.

7.3.3 Dimensions 4 y 5

Finally we realize a superficial study of the individuals map and the variables obtained from the dimensions 4 and 5.



Where we can observe the following group of interrelated variables:

- Accuage - year

In relation to the axes chosen for the study (1, 2) we see now that the variables that did not have a significant representation as the age of the victim, the age of the accused and the date of the crime, now go on to play an important role. Otherwise, all variables related to the characteristics of the population (density, level of education, wealth, percentage of races, number of men and women) to stop having such a remarkable significance.

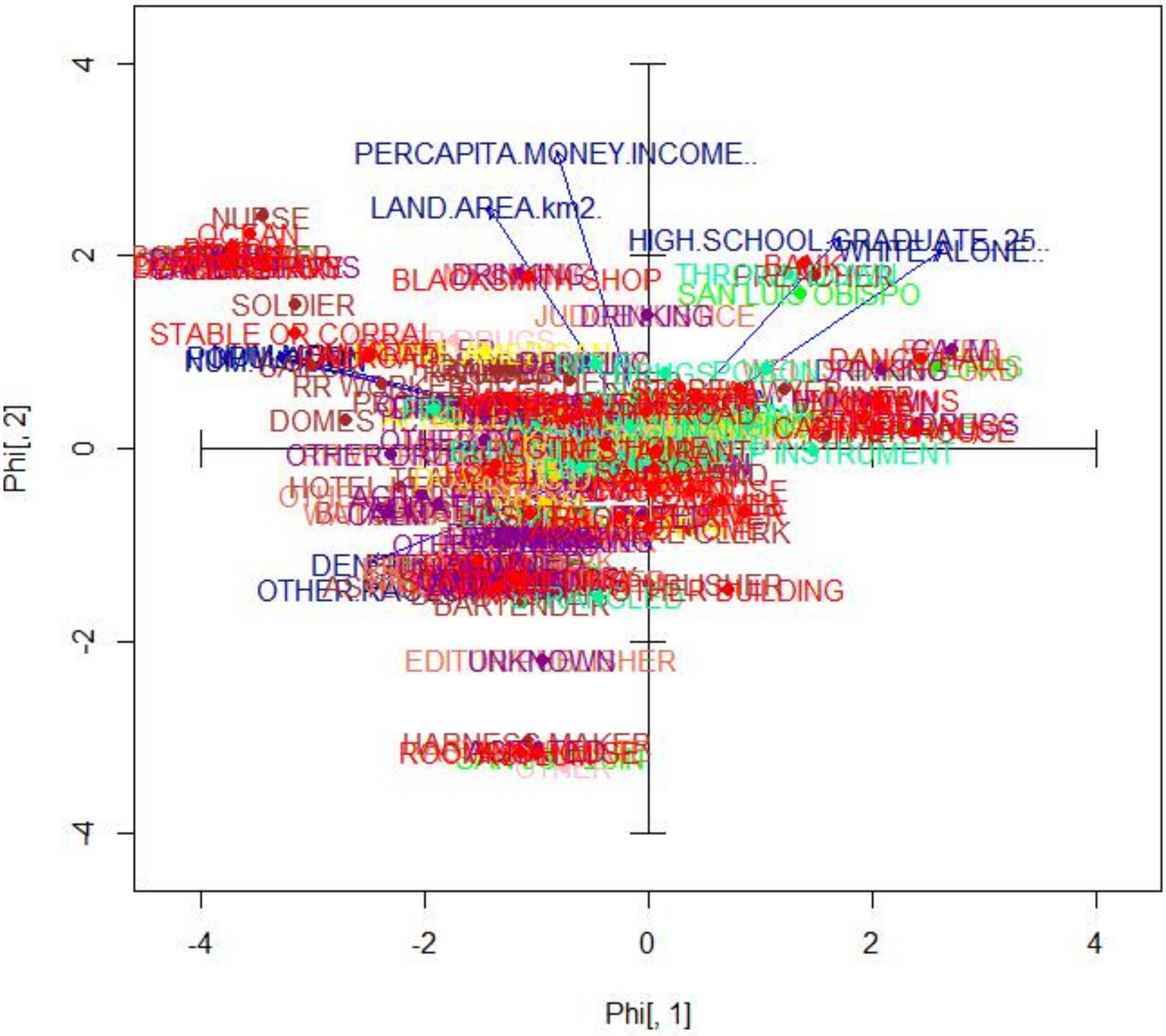
7.3.4 Conclusion

After analyzing the representation of variables according to the chosen axis, we can determine that the best axes for our study are 1 and 2. They are the ones that provide more meaning to the difference between the percentages of races and characteristics of the population. Although, to future studies for example, the axes 4 and 5 would be useful for a study on the distribution of crimes by date and age of victims and accused. Also the axes 3 and 4 could be used for a study of the crimes according to the date and the characteristics of wealth, education and population density.

7.4 Common projection of numerical variables and qualitative modalities

7.4.1 Dimensions 1 y 2

Next we proceed to project the qualitative variables on quantitative variables that we have seen before in order to try to extract as much information as possible and observe how they relate.

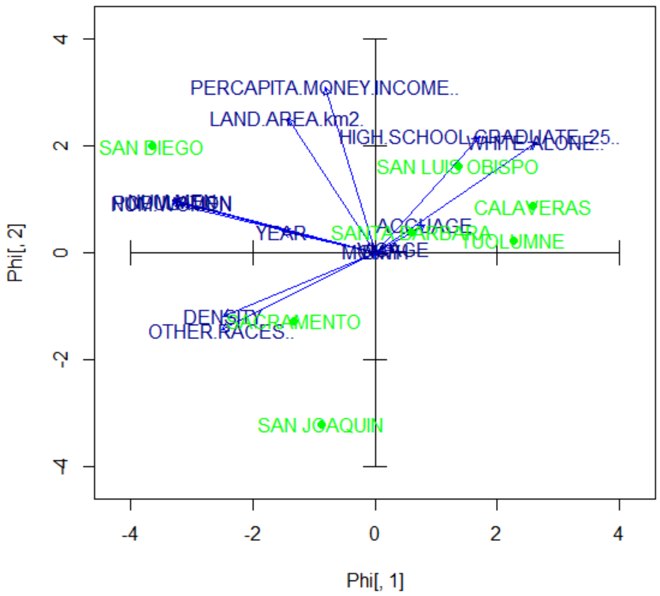


By having a large number of qualitative variables and a great variability for many of these we obtain a figure difficult to interpret.

So we decided to divide it into different graphs with the same objective but a smaller number of qualitative variables simultaneously. This way we hope to obtain less overwhelmed and more easily interpretable graphics.

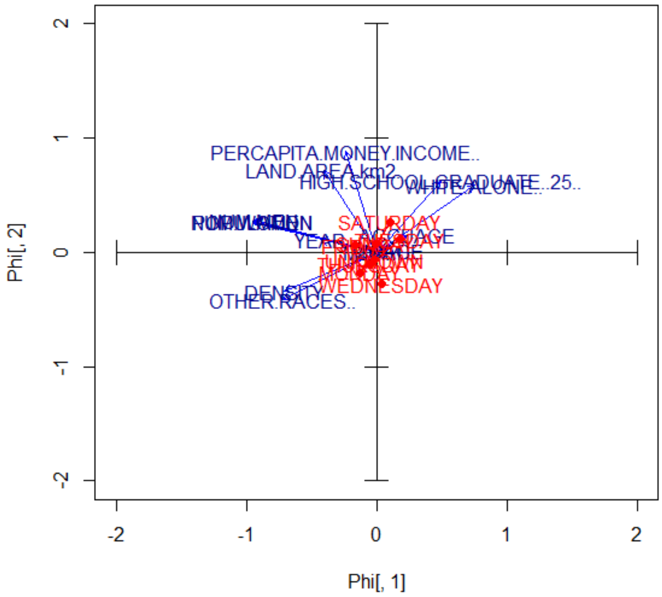
The following for each qualitative variable we will represent the graph alongside quantitative variables.

7.4.1.1 COUNTY



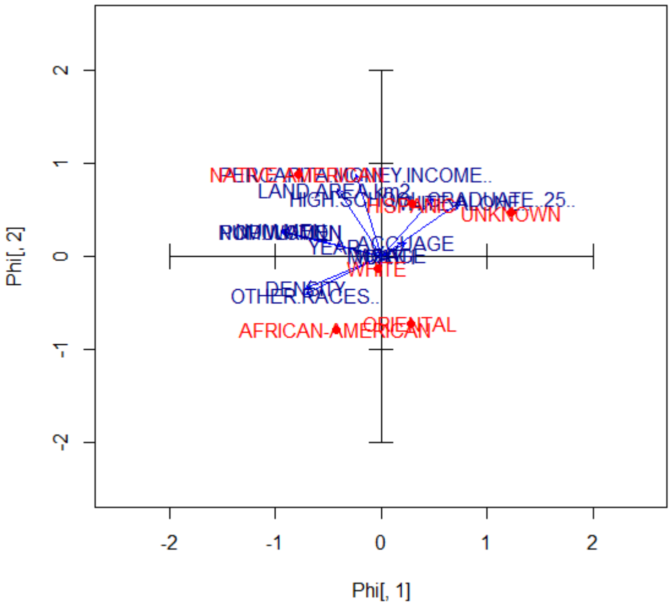
It is observed that counties like San Joaquin and Sacramento have a higher population density and greater diversity of races that San Diego and San Luis Obispo (Dimension 2). On the other hand we note that San Diego has a larger population and surface than Tuolumne and Calaveras.

7.4.1.2 WEEKDAY



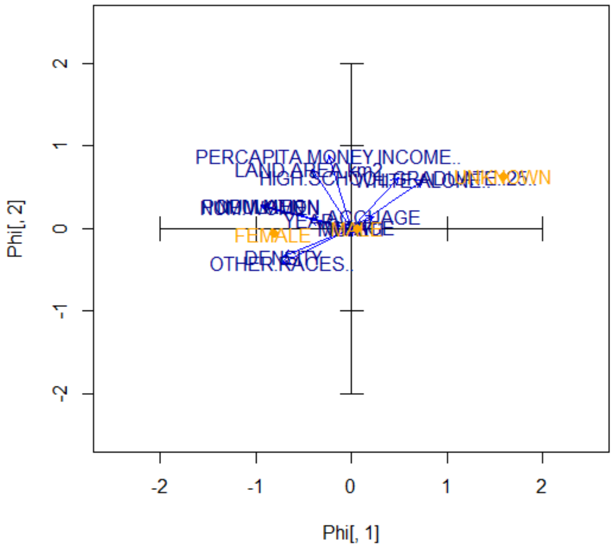
It is observed that the day of the week that the event has happened shows no relation with the other variables.

7.4.1.3 VICRACE



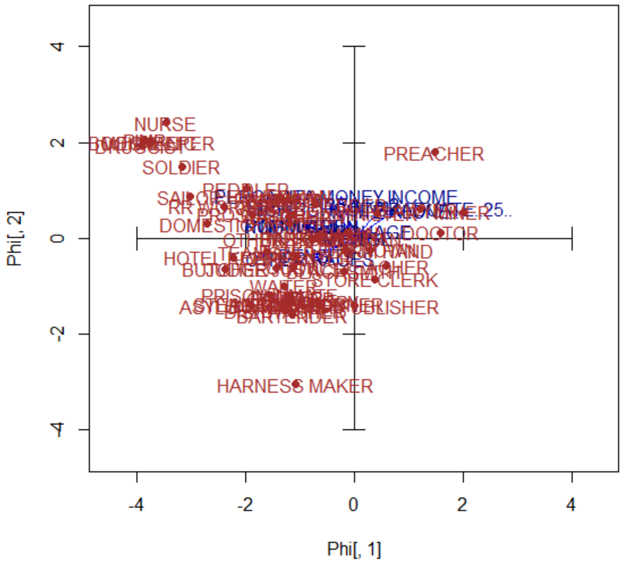
It is observed that the races African-American and Oriental are related to large population densities, that the white race is evenly distributed to the rest of variables and Hispanic and Native American races are linked with higher income, studies and surface.

7.4.1.4 VICSEX



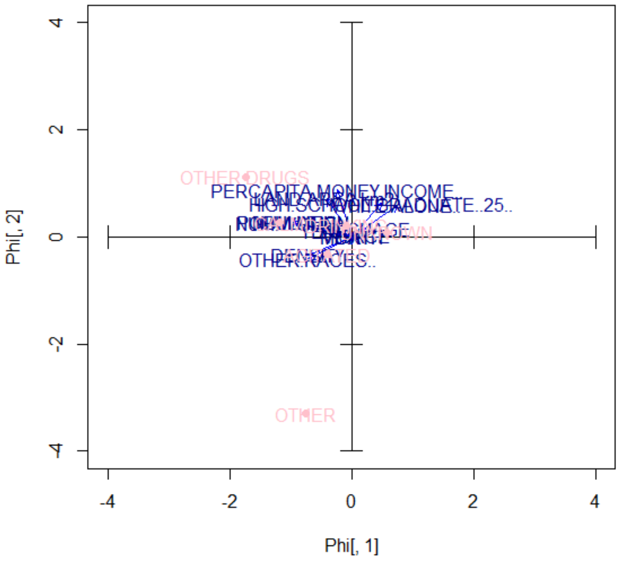
It is observed that female victims is related to populations with a lower percentage of higher education, greater amount of population of other races and large population densities. While the male is uniformly distributed for all variables.

7.4.1.5 VICOCCUP



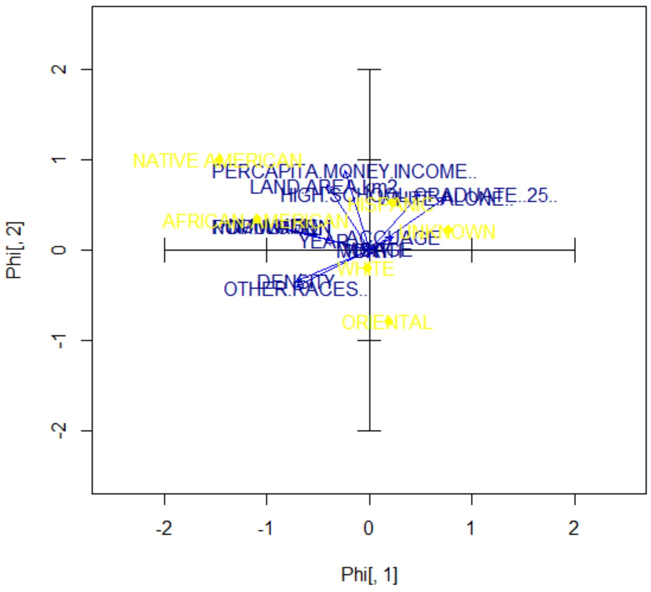
It is observed that the less qualified professions are more related to lower percentage of higher education as well as lower incomes.

7.4.1.6 VICCOND



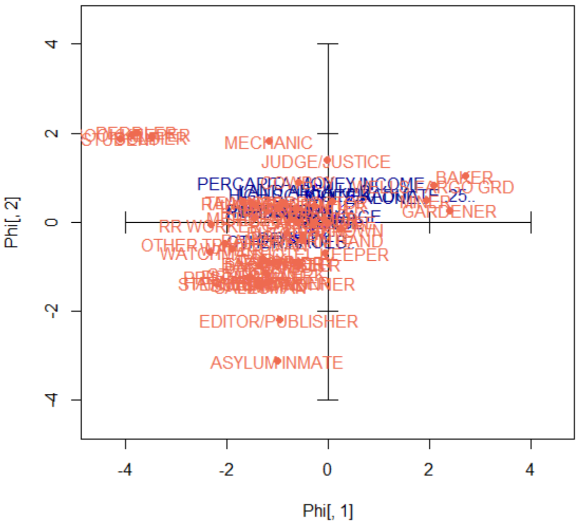
It is not observed a great relationship between the variables and the condition of the victim. Even though we can see that individuals with higher incomes and studies behold a greater relation with other kind of drugs.

7.4.1.7 ACCURACE



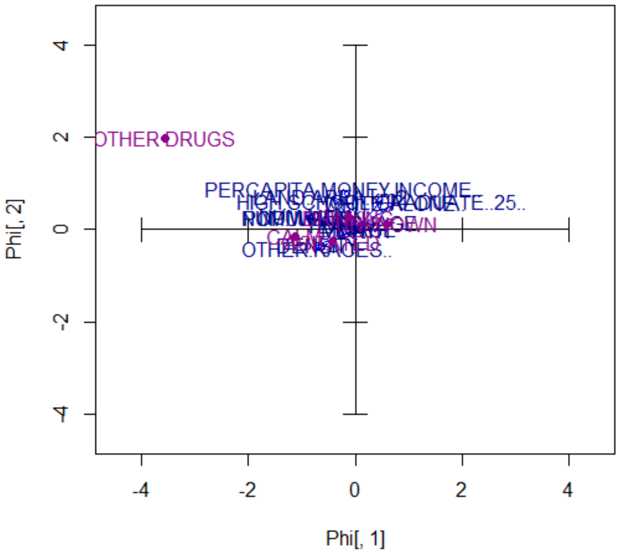
It is observed that for the race of the accused in the case of African-American and Native American races are related to greater surfaces of land and higher densities, while the white race is evenly distributed all variables.

7.4.1.8 ACCUOCCU



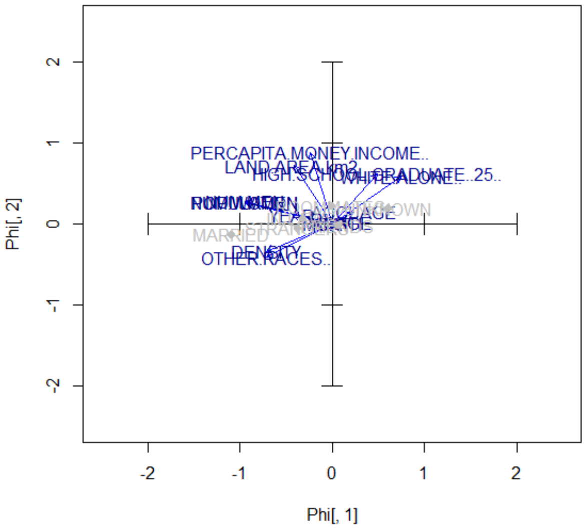
It is observed that the less qualified professions are more related to lower percentage of higher education as well as lower incomes.

7.4.1.9 ACCUCOND



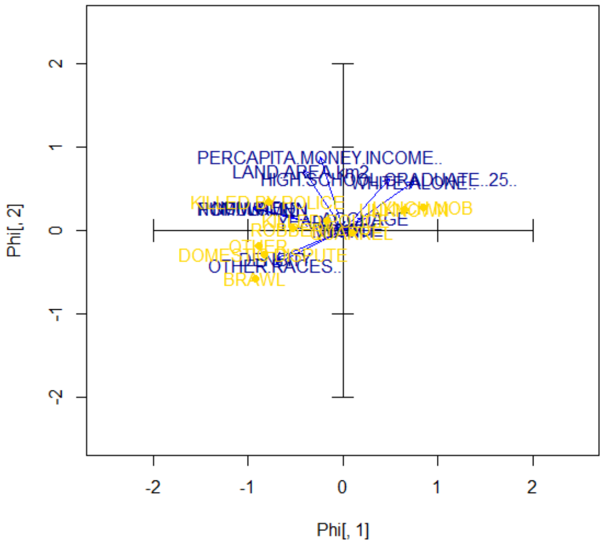
It is observed that for individuals with less education and less income the condition in which they are usually found is agitated, whereas for individuals with higher levels of education and income main condition is usually drunk or other drugs.

7.4.1.10 RELATION



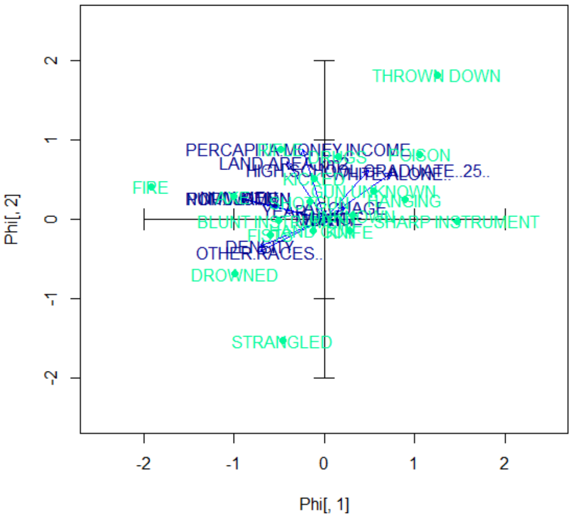
There is no great relation observed at the type of relation between the individual the variables is observed. Worth noting that married people belong to populations with high densities and a higher percentage of other races.

7.4.1.11 CAUSE



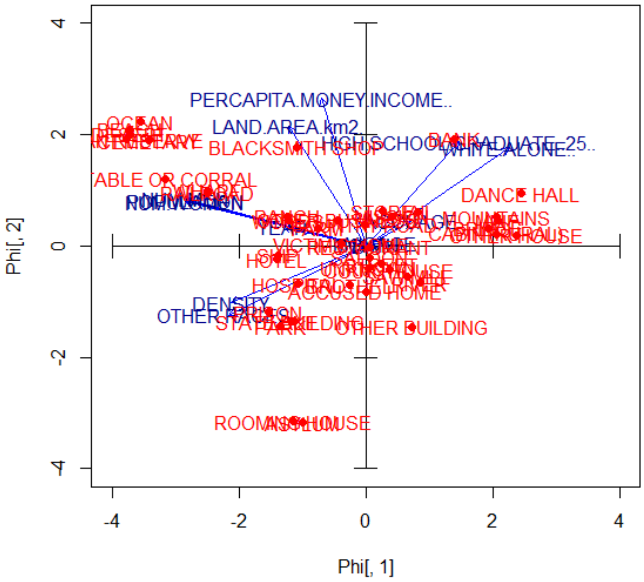
We note that offenses such as fighting, theft, domestic fights or killings by police are more likely to occur in counties with higher population density, percentage of low education and low income. While for other crime related to the mafia and the police tend to occur in counties with higher incomes and higher rates of education.

7.4.1.12 WEAPON



It is observed that in populations with higher population density, lower incomes and less education they are more common the less premeditated crimes and in a simple way(strangled), while with the increasing percentage of income and studies, crimes tend to be more premeditated and complex (Poison, guns).

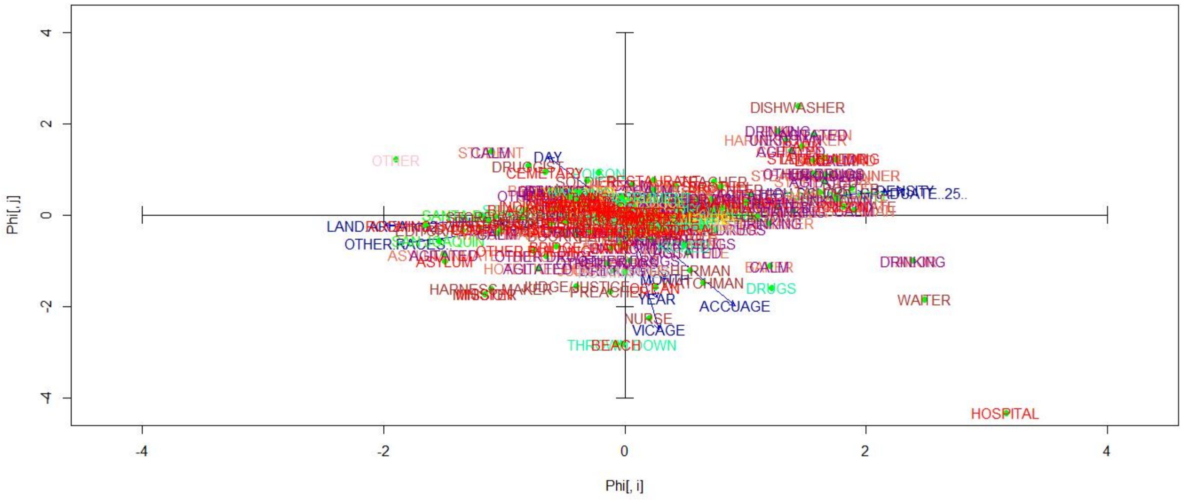
7.4.1.13 LOCATION



We note that the variables greatest number of incomes and higher percentage of studies are related with “selected” places, as it could be places like banks and dance floors, while lower income, less educated and higher population densities are associated with less "selected" locations such as farms, parks, etc.

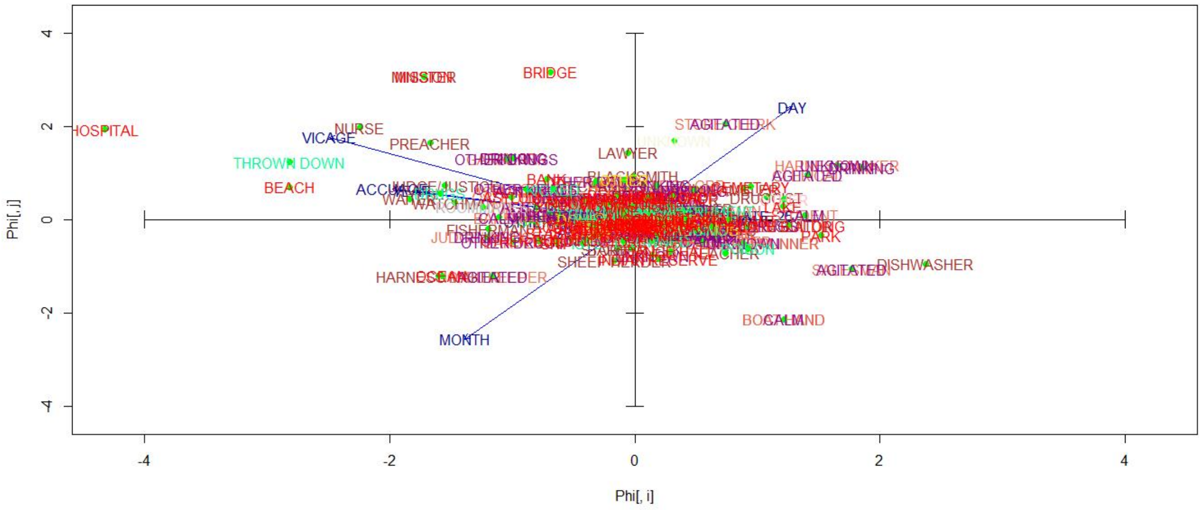
7.4.2 Dimensions 3,4

Next we proceed to project the qualitative variables on quantitative variables that we have seen in dimensions 3 and 4. We observe that as in the above section that holding too many variables overloads the map thus making it quite difficult to interpret, and to improve the interpretability of the graphics would be optimal to divide the plane according to different variables.



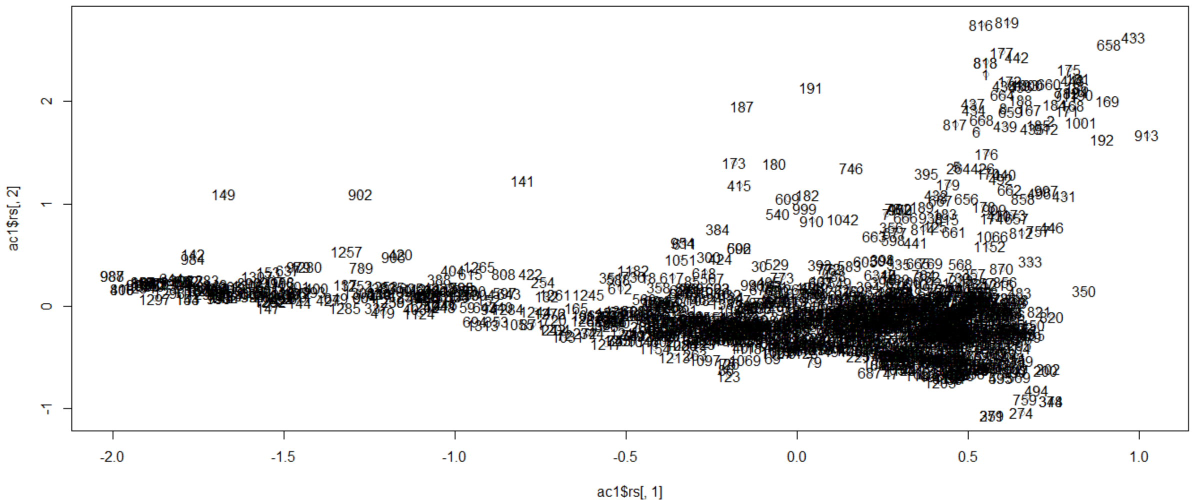
7.4.3 Dimensions 4,5

Next we proceed to project the qualitative variables on quantitative variables that we have seen in dimensions 4 and 5. We observe that as in the above section that holding too many variables overloads the map thus making it quite difficult to interpret, and to improve the interpretability of the graphics would be optimal to divide the plane according to different variables.

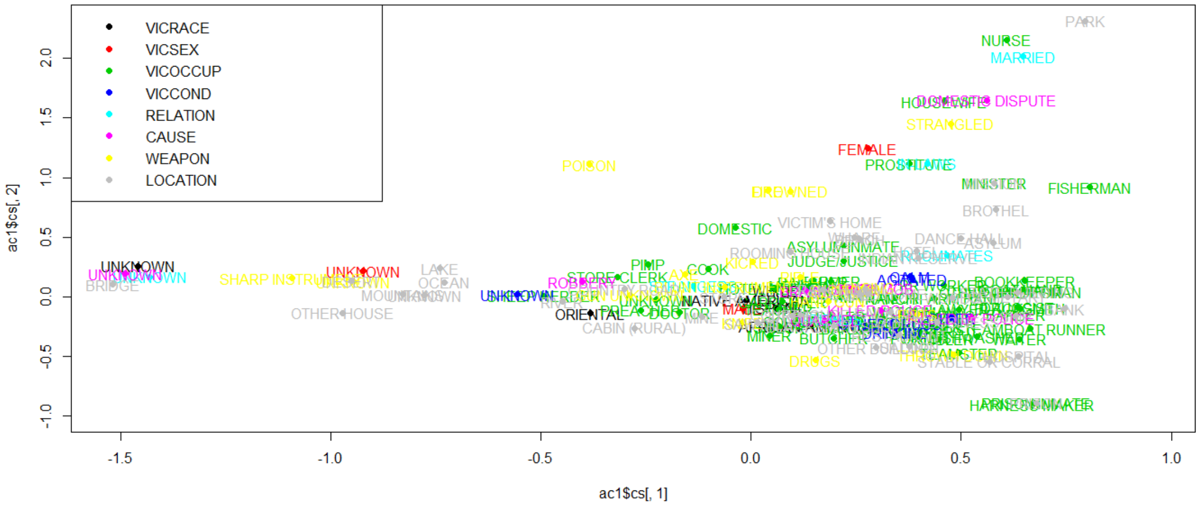


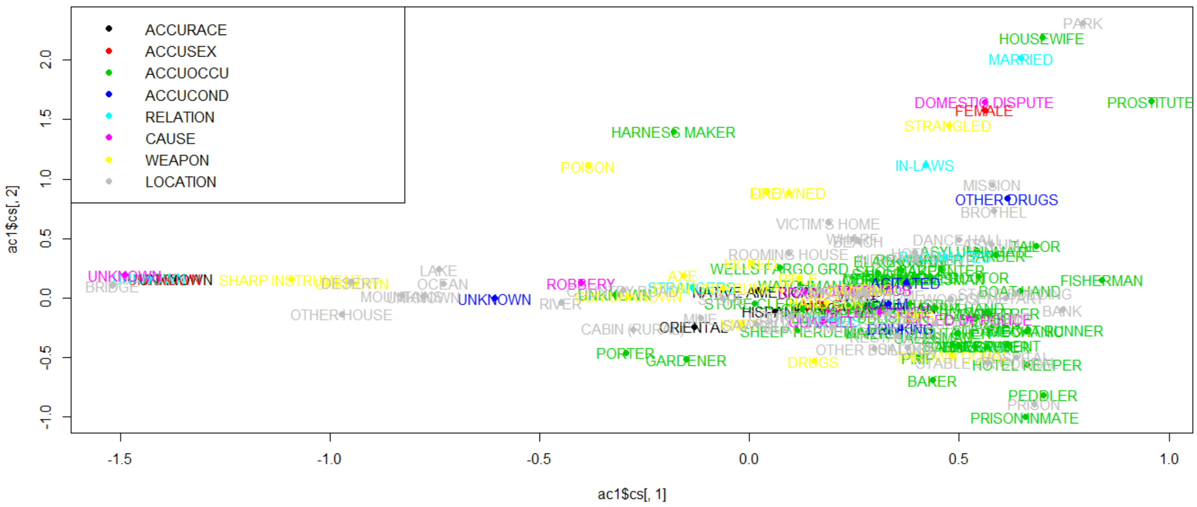
7.5 MCA

Next, after have the ACM applied on our data set , we attach the cloud of individuals obtained.



We proceed to relate the variables through ACM, with this purpose we have differentiated between victims and accused, otherwise we get a too overloaded and very difficult to interpret graph.



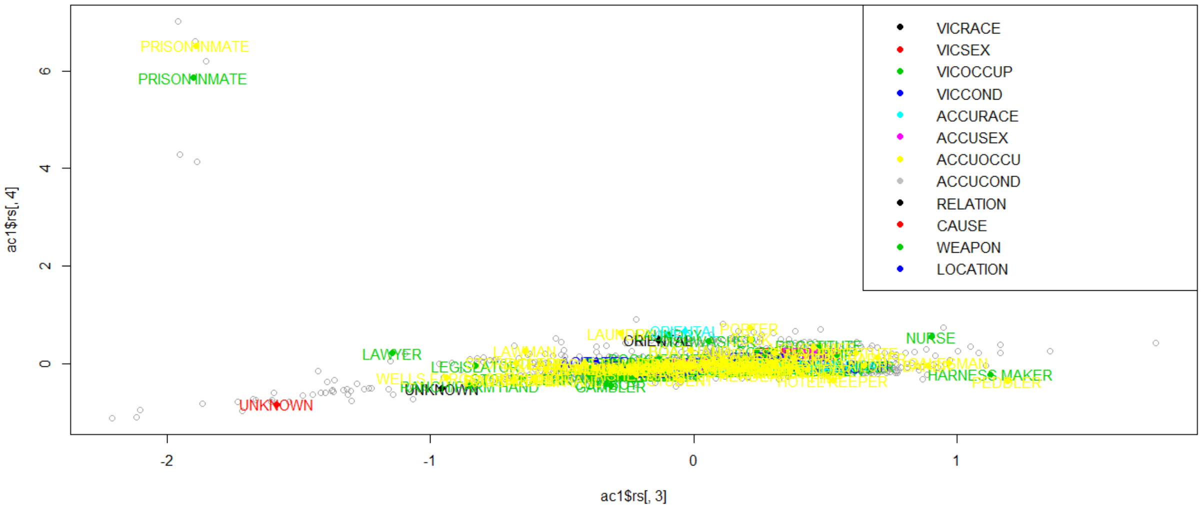


From the above figures we get different conclusions:

- We observed relations between female sex, prostitute profession and locations such as parks or homes. That is, women prostitutes are often victims who are killed in parks or homes. The same goes for women who are housewives, who are murdered in homes and usually strangled.
- We observe that for those variables with unknown values the most likely related location are the bridges, that can match suicides on bridges. That is, for those people who have thrown off a bridge, at the time to write the incident report possibly could not fill all the data.
- We observe that sea-related professions as (Sharks intructor, etc) are often victims in marine locations such as lakes and oceans.
- We observe that when the place is the prison of the murder victims and accused are prisoners.

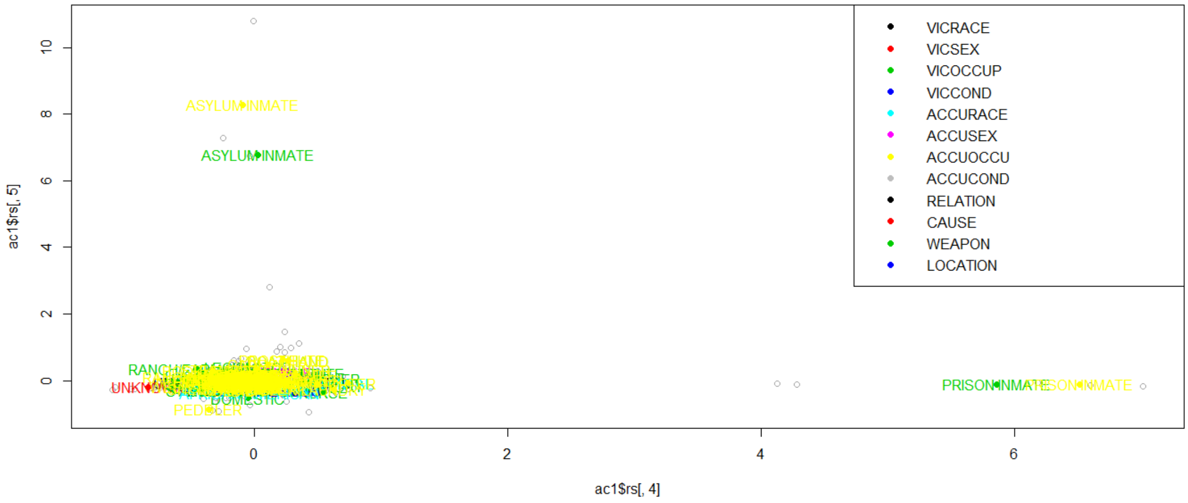
7.5.1 MCA dim 3, 4

We have also performed these graphics on the dimensions 3 and 4. As we got a too overloaded and less informative graphics we preferred to discard them.



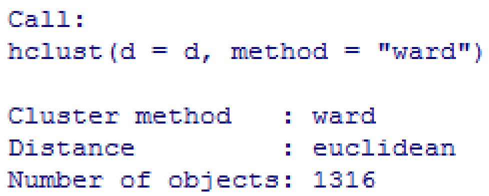
7.5.2 MCA 4,5

We have also performed these graphics on the dimensions 4 and 5. As we got a too overloaded and less informative graphics we preferred to discard them.

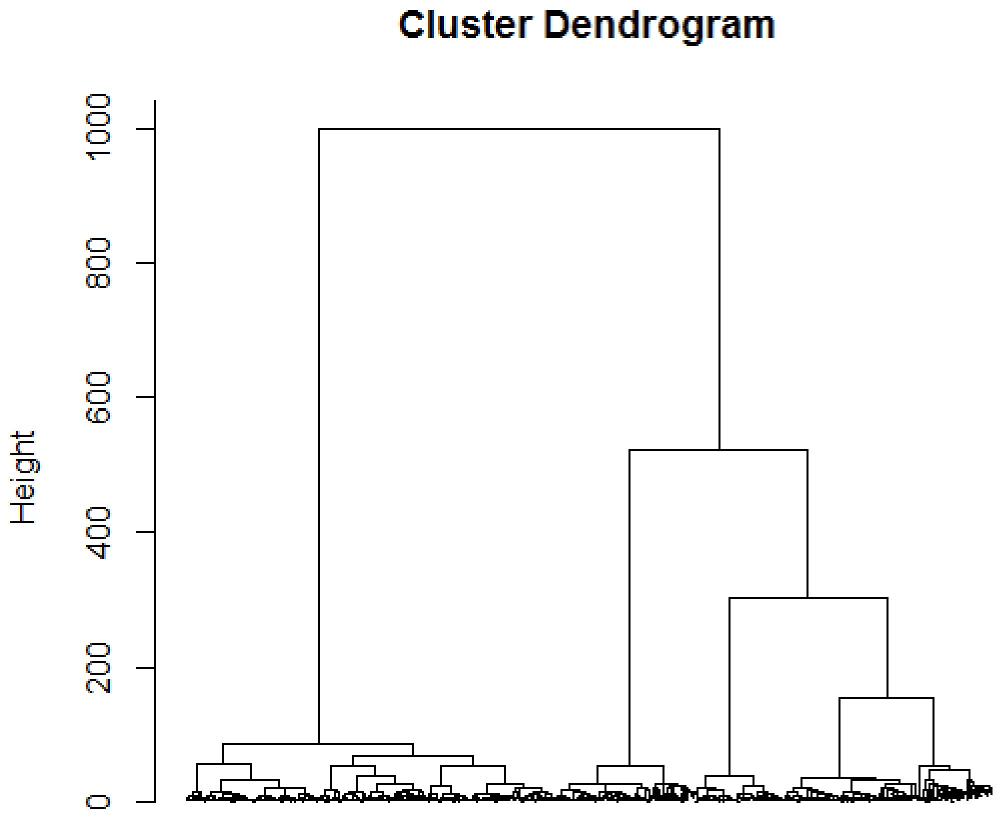


8. Hierarchical Clustering analysis

In this section we shall analyze which are the different clusters that are formed from our data set. We will use the ward method to find such clusters.

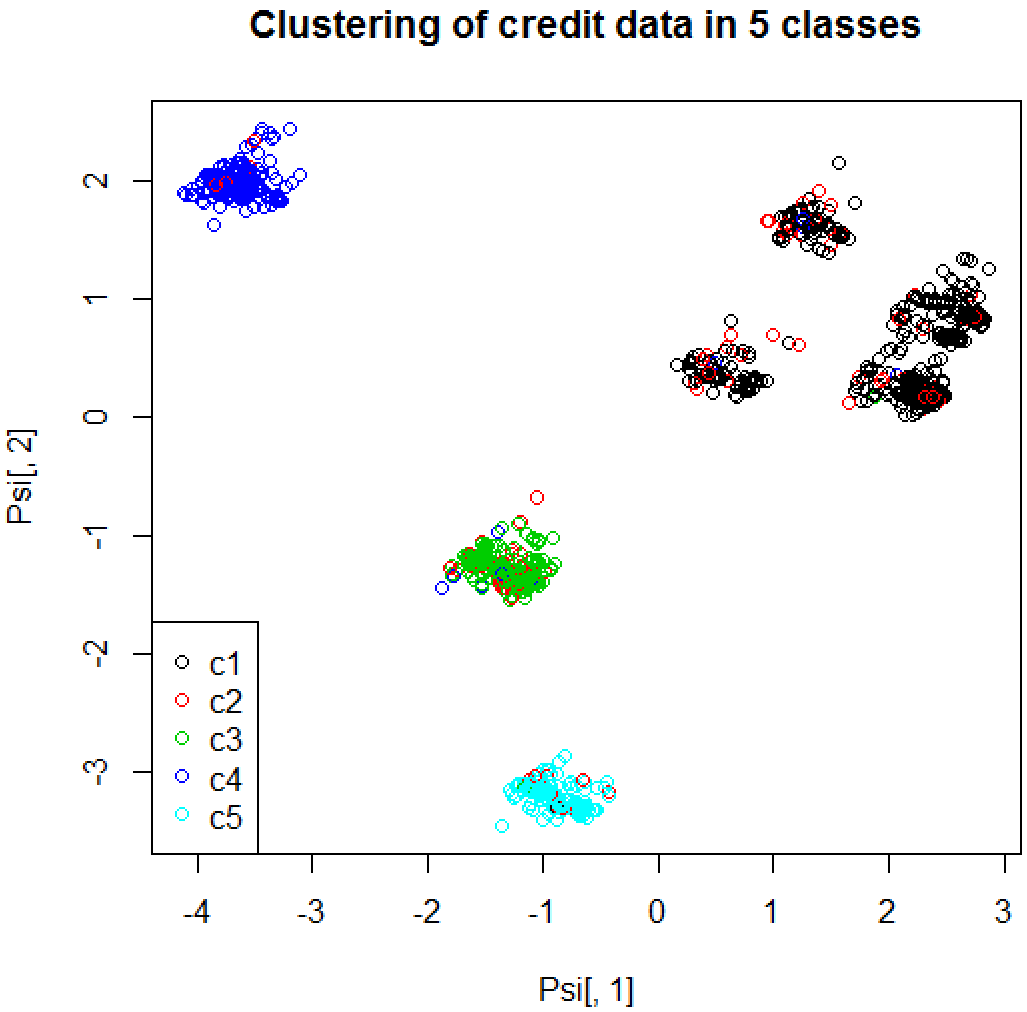


```
Call:
hclust(d = d, method = "ward")

Cluster method   : ward
Distance         : euclidean
Number of objects: 1316
```



As we can see from the above graphic, our data has many levels to make various clusters. We decided to cut the tree making 5 different clusters. Performing that cut, we obtain the clusters that are represented in the next graphic.

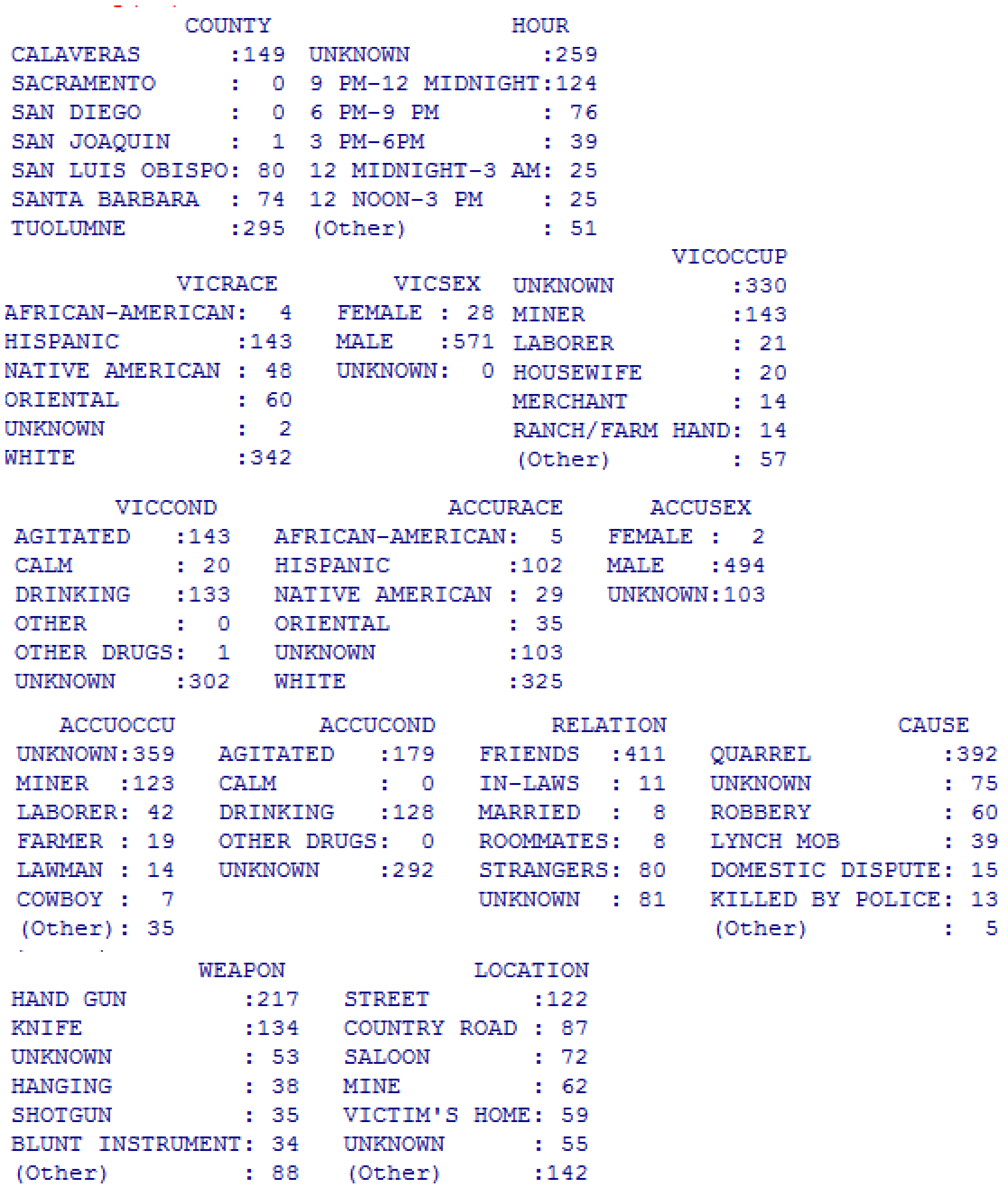
9. Profiling of clusters



As we can see in the previous figure, it is possible to distinguish 4 out of 5 clusters in a very easy way in the cloud of individuals that distinguish them mainly by the county in which the murders occurred. The other cluster is divided in the different counties.

Having reached this point, we will analyze the profiles of each of the clusters from the data of individuals that are part of each one of them. From these data we will see what are the characteristics that distinguish the 5 profiles of the clusters.

Cluster 1

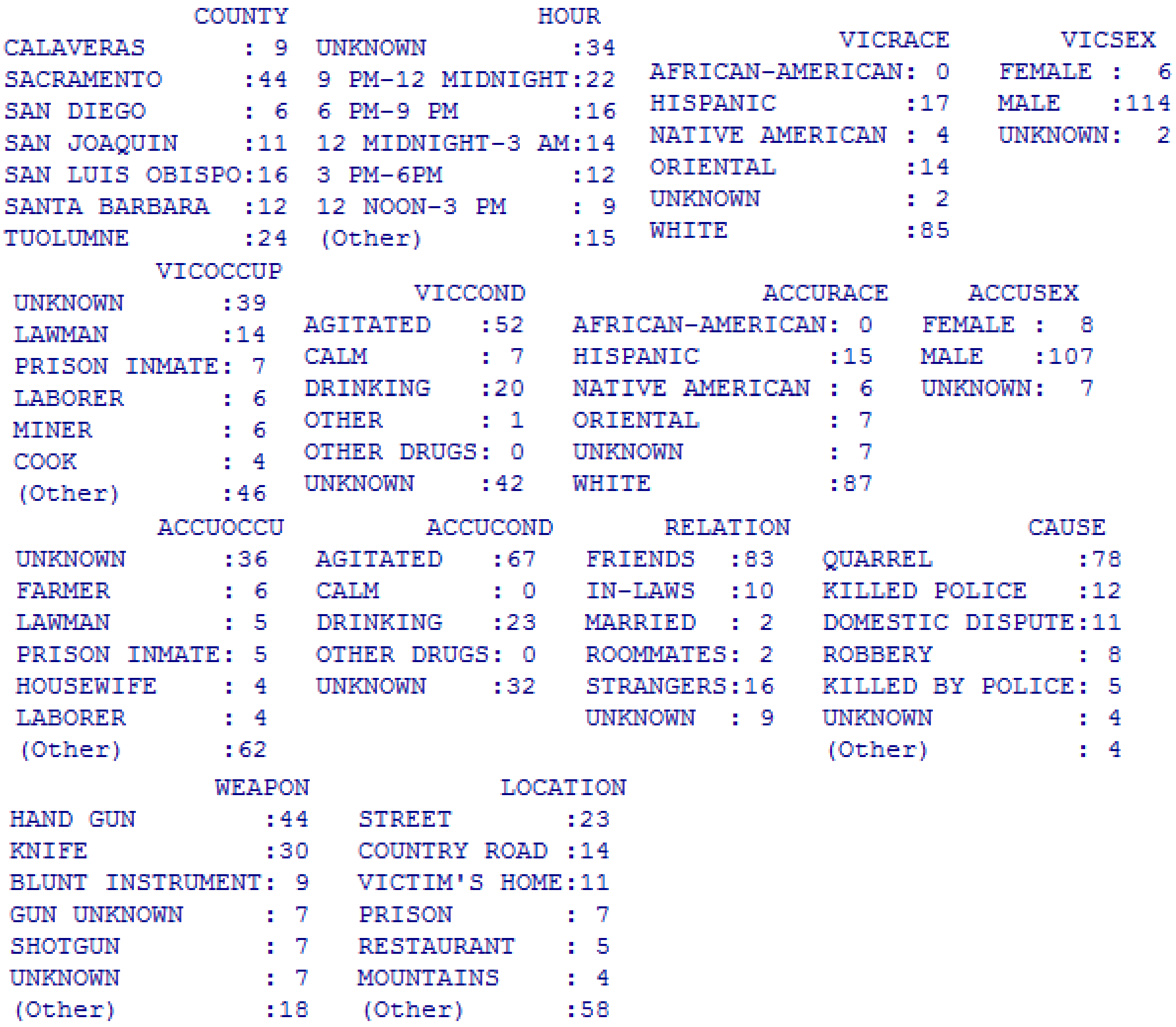


```
            COUNTY                      HOUR
CALAVERAS         :149  UNKNOWN            :259
SACRAMENTO        :  0  9 PM-12 MIDNIGHT:124
SAN DIEGO         :  0  6 PM-9 PM         : 76
SAN JOAQUIN       :  1  3 PM-6PM          : 39
SAN LUIS OBISPO: 80  12 MIDNIGHT-3 AM: 25
SANTA BARBARA   : 74  12 NOON-3 PM      : 25
TUOLUMNE          :295  (Other)           : 51
                                                      VICOCCUP
            VICRACE           VICSEX    UNKNOWN           :330
AFRICAN-AMERICAN:   4   FEMALE : 28  MINER             :143
HISPANIC            :143   MALE    :571  LABORER           : 21
NATIVE AMERICAN : 48   UNKNOWN:   0  HOUSEWIFE         : 20
ORIENTAL            : 60                  MERCHANT          : 14
UNKNOWN             :  2                  RANCH/FARM HAND: 14
WHITE               :342                  (Other)           : 57

        VICCOND                 ACCURACE           ACCUSEX
AGITATED    :143   AFRICAN-AMERICAN:   5   FEMALE :   2
CALM        : 20   HISPANIC            :102   MALE    :494
DRINKING    :133   NATIVE AMERICAN : 29   UNKNOWN:103
OTHER       :  0   ORIENTAL            : 35
OTHER DRUGS:  1   UNKNOWN             :103
UNKNOWN     :302   WHITE               :325

   ACCUOCCU          ACCUCOND          RELATION              CAUSE
UNKNOWN:359   AGITATED    :179   FRIENDS   :411   QUARREL           :392
MINER   :123   CALM        :  0   IN-LAWS   : 11   UNKNOWN           : 75
LABORER: 42   DRINKING    :128   MARRIED   :  8   ROBBERY           : 60
FARMER : 19   OTHER DRUGS:  0   ROOMMATES:  8   LYNCH MOB         : 39
LAWMAN : 14   UNKNOWN     :292   STRANGERS: 80   DOMESTIC DISPUTE: 15
COWBOY :  7                      UNKNOWN   : 81   KILLED BY POLICE: 13
(Other): 35                                         (Other)           :  5

              WEAPON                LOCATION
HAND GUN          :217   STREET          :122
KNIFE             :134   COUNTRY ROAD : 87
UNKNOWN           : 53   SALOON          : 72
HANGING           : 38   MINE            : 62
SHOTGUN           : 35   VICTIM'S HOME: 59
BLUNT INSTRUMENT: 34   UNKNOWN         : 55
(Other)           : 88   (Other)         :142
```

From the first cluster we can highlight the following characteristics:

- It is a group represented by individuals from 4 counties (San Luis Obispo, Santa Barbara, Calaveras y Tulumne). We can say that those are counties with mining exploitations because as we can see that miner is the main occupation of victim and the accused.
- Crimes occur mainly at night.
- The race of the victim and the accused is mainly white man. Also, most of the predominant sex of victim and accused is man.
- Finally we can say that in most cases the relation between victim and accused is friendship.

Cluster 2



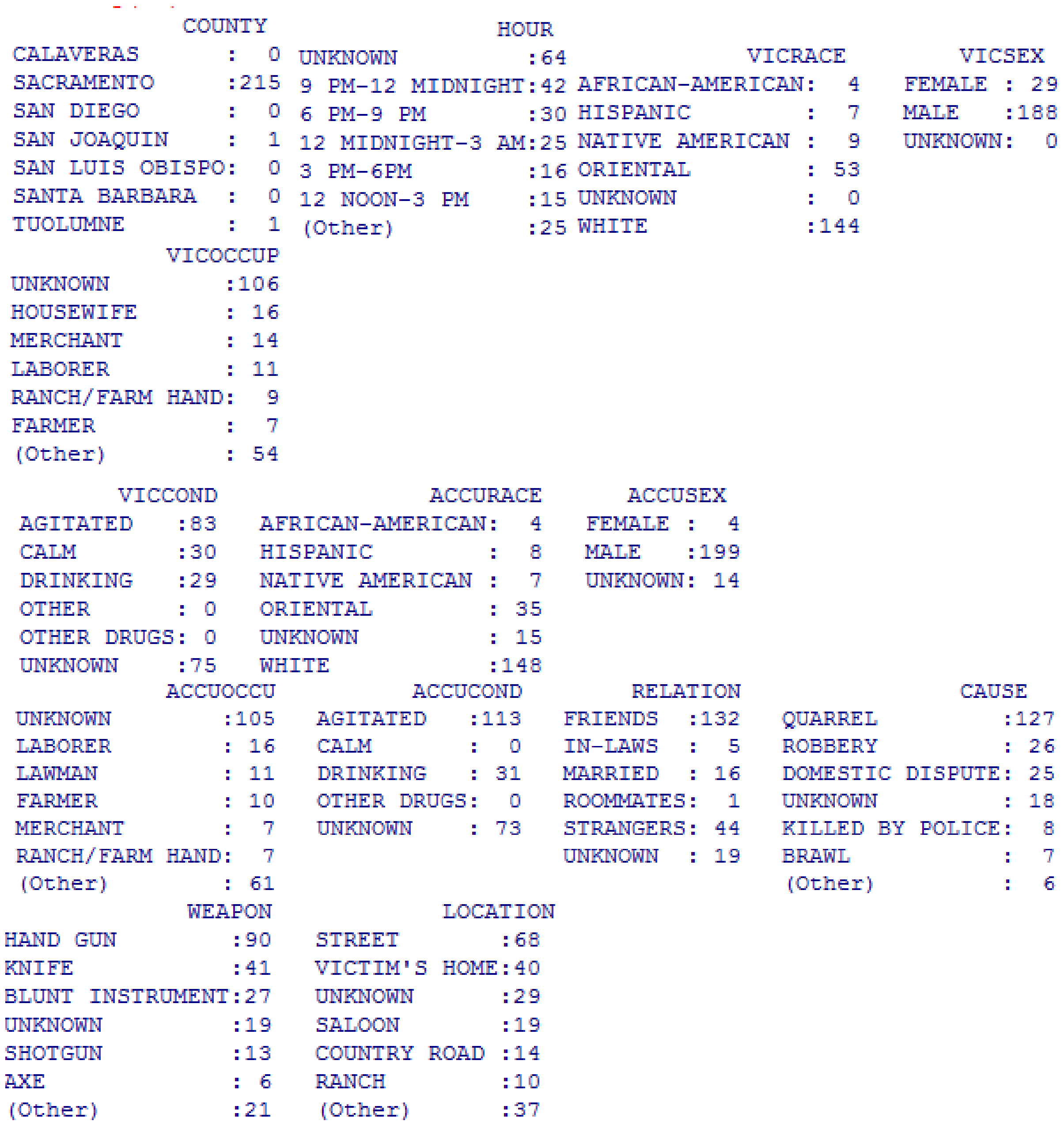
```
            COUNTY                     HOUR
CALAVERAS      : 9  UNKNOWN          :34                VICRACE          VICSEX
SACRAMENTO     :44  9 PM-12 MIDNIGHT:22  AFRICAN-AMERICAN: 0    FEMALE :  6
SAN DIEGO      : 6  6 PM-9 PM        :16  HISPANIC         :17   MALE   :114
SAN JOAQUIN    :11  12 MIDNIGHT-3 AM:14  NATIVE AMERICAN : 4    UNKNOWN:  2
SAN LUIS OBISPO:16  3 PM-6PM         :12  ORIENTAL         :14
SANTA BARBARA  :12  12 NOON-3 PM     : 9  UNKNOWN          : 2
TUOLUMNE       :24  (Other)          :15  WHITE            :85
          VICOCCUP
 UNKNOWN       :39          VICCOND                 ACCURACE         ACCUSEX
 LAWMAN        :14   AGITATED    :52   AFRICAN-AMERICAN: 0   FEMALE :  8
 PRISON INMATE: 7    CALM        : 7   HISPANIC        :15   MALE   :107
 LABORER       : 6   DRINKING    :20   NATIVE AMERICAN : 6   UNKNOWN:  7
 MINER         : 6   OTHER       : 1   ORIENTAL        : 7
 COOK          : 4   OTHER DRUGS: 0    UNKNOWN         : 7
 (Other)       :46   UNKNOWN     :42   WHITE           :87
          ACCUOCCU          ACCUCOND          RELATION              CAUSE
 UNKNOWN       :36   AGITATED    :67   FRIENDS  :83   QUARREL          :78
 FARMER        : 6   CALM        : 0   IN-LAWS  :10   KILLED POLICE    :12
 LAWMAN        : 5   DRINKING    :23   MARRIED  : 2   DOMESTIC DISPUTE:11
 PRISON INMATE: 5    OTHER DRUGS: 0    ROOMMATES: 2   ROBBERY          : 8
 HOUSEWIFE     : 4   UNKNOWN     :32   STRANGERS:16   KILLED BY POLICE: 5
 LABORER       : 4                     UNKNOWN  : 9   UNKNOWN          : 4
 (Other)       :62                                    (Other)          : 4
            WEAPON              LOCATION
 HAND GUN         :44   STREET         :23
 KNIFE            :30   COUNTRY ROAD :14
 BLUNT INSTRUMENT: 9    VICTIM'S HOME:11
 GUN UNKNOWN      : 7   PRISON         : 7
 SHOTGUN          : 7   RESTAURANT     : 5
 UNKNOWN          : 7   MOUNTAINS      : 4
 (Other)          :18   (Other)        :58
```

From the second cluster we can highlight the following characteristics:

- It is a group represented by individuals from all counties.
- The race of the victim and the accused are in most cases white. Also in most cases are prisoners or law enforcement.
- The relation between the victim and the accused is friendship or prison mates, mainly.
- The homicide begins in a fight and ends with the victim on the street or a highway of the county, but also at home or in jail.
- Homicide is committed with firearms or knives.

  In this group we would distinguish homicides in prisons between jail mates, cops killed in fights, in theft or in the hands of the police.

Cluster 3



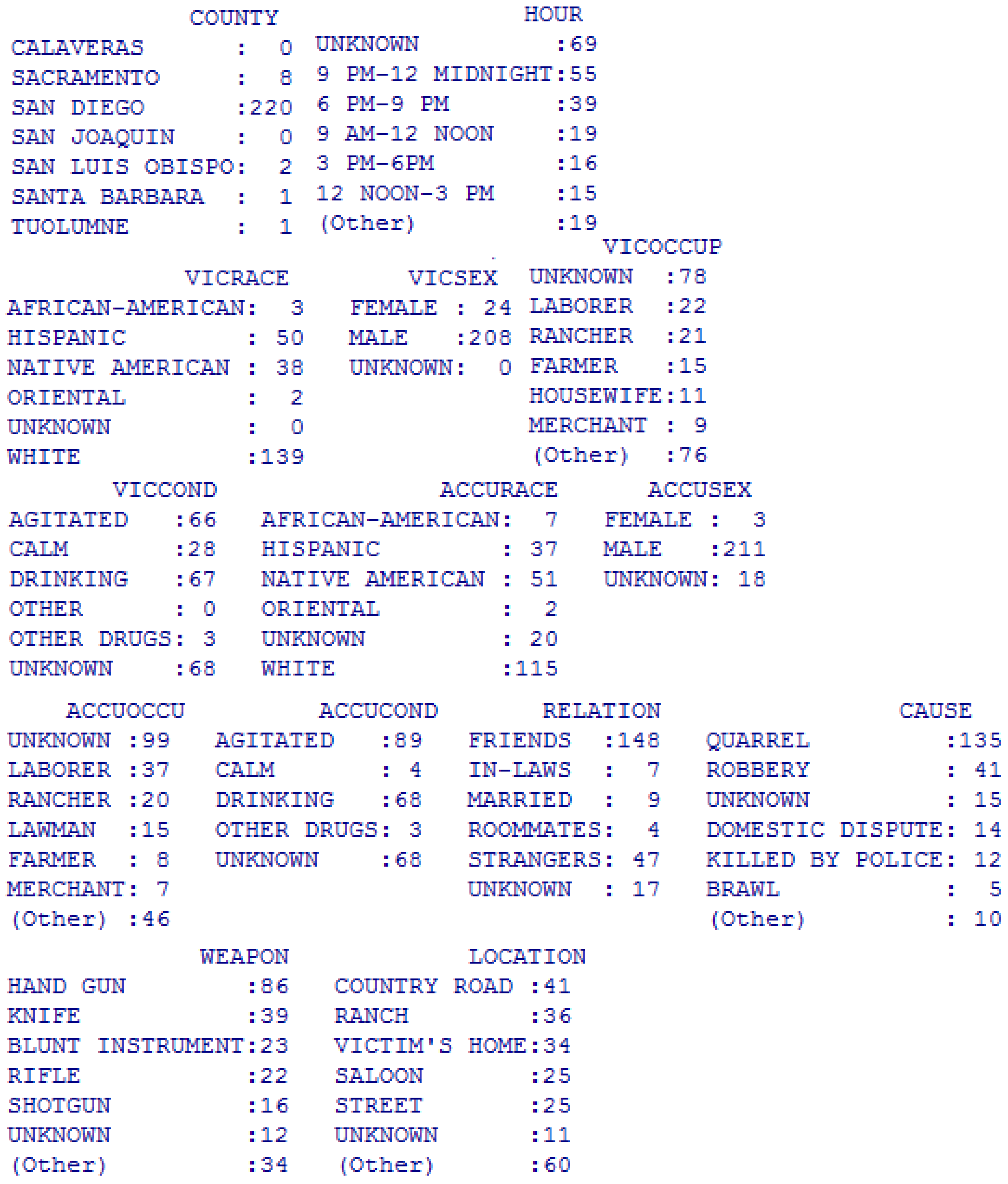
```
              COUNTY                      HOUR
CALAVERAS      :  0 UNKNOWN          :64           VICRACE            VICSEX
SACRAMENTO     :215 9 PM-12 MIDNIGHT:42 AFRICAN-AMERICAN:  4   FEMALE : 29
SAN DIEGO      :  0 6 PM-9 PM        :30 HISPANIC         :  7   MALE   :188
SAN JOAQUIN    :  1 12 MIDNIGHT-3 AM:25 NATIVE AMERICAN :  9   UNKNOWN:  0
SAN LUIS OBISPO:  0 3 PM-6PM         :16 ORIENTAL         : 53
SANTA BARBARA  :  0 12 NOON-3 PM     :15 UNKNOWN          :  0
TUOLUMNE       :  1 (Other)          :25 WHITE            :144
           VICOCCUP
UNKNOWN        :106
HOUSEWIFE      : 16
MERCHANT       : 14
LABORER        : 11
RANCH/FARM HAND:  9
FARMER         :  7
(Other)        : 54

       VICCOND                 ACCURACE          ACCUSEX
 AGITATED    :83  AFRICAN-AMERICAN:  4   FEMALE :  4
 CALM        :30  HISPANIC         :  8   MALE   :199
 DRINKING    :29  NATIVE AMERICAN :  7   UNKNOWN: 14
 OTHER       : 0  ORIENTAL         : 35
 OTHER DRUGS: 0  UNKNOWN          : 15
 UNKNOWN     :75  WHITE            :148
           ACCUOCCU          ACCUCOND          RELATION                 CAUSE
 UNKNOWN        :105   AGITATED    :113   FRIENDS  :132   QUARREL          :127
 LABORER        : 16   CALM        :  0   IN-LAWS  :  5   ROBBERY          : 26
 LAWMAN         : 11   DRINKING    : 31   MARRIED  : 16   DOMESTIC DISPUTE: 25
 FARMER         : 10   OTHER DRUGS:  0   ROOMMATES:  1   UNKNOWN          : 18
 MERCHANT       :  7   UNKNOWN     : 73   STRANGERS: 44   KILLED BY POLICE:  8
 RANCH/FARM HAND:  7                      UNKNOWN  : 19   BRAWL            :  7
 (Other)        : 61                                      (Other)          :  6
             WEAPON            LOCATION
HAND GUN         :90   STREET       :68
KNIFE            :41   VICTIM'S HOME:40
BLUNT INSTRUMENT:27   UNKNOWN      :29
UNKNOWN          :19   SALOON       :19
SHOTGUN          :13   COUNTRY ROAD :14
AXE              : 6   RANCH        :10
(Other)          :21   (Other)      :37
```

From the third cluster we can highlight the following characteristics:

- It is a group represented by individuals of Sacramento.
- The victims are mostly white and male. But we note that the most represented occupation is housewife and workers.
- Those accused are mostly white men and their main occupation is worker.
- Homicide in most cases begins in a fight, theft or domestic dispute and ends with the victim on the street or at home.
  - Homicide is committed with firearms, knives, or any blunt object like hammers or bats.

- From within this group we would distinguish the homicides by domestic disputes, robberies or burglaries, besides fights..

Cluster 4

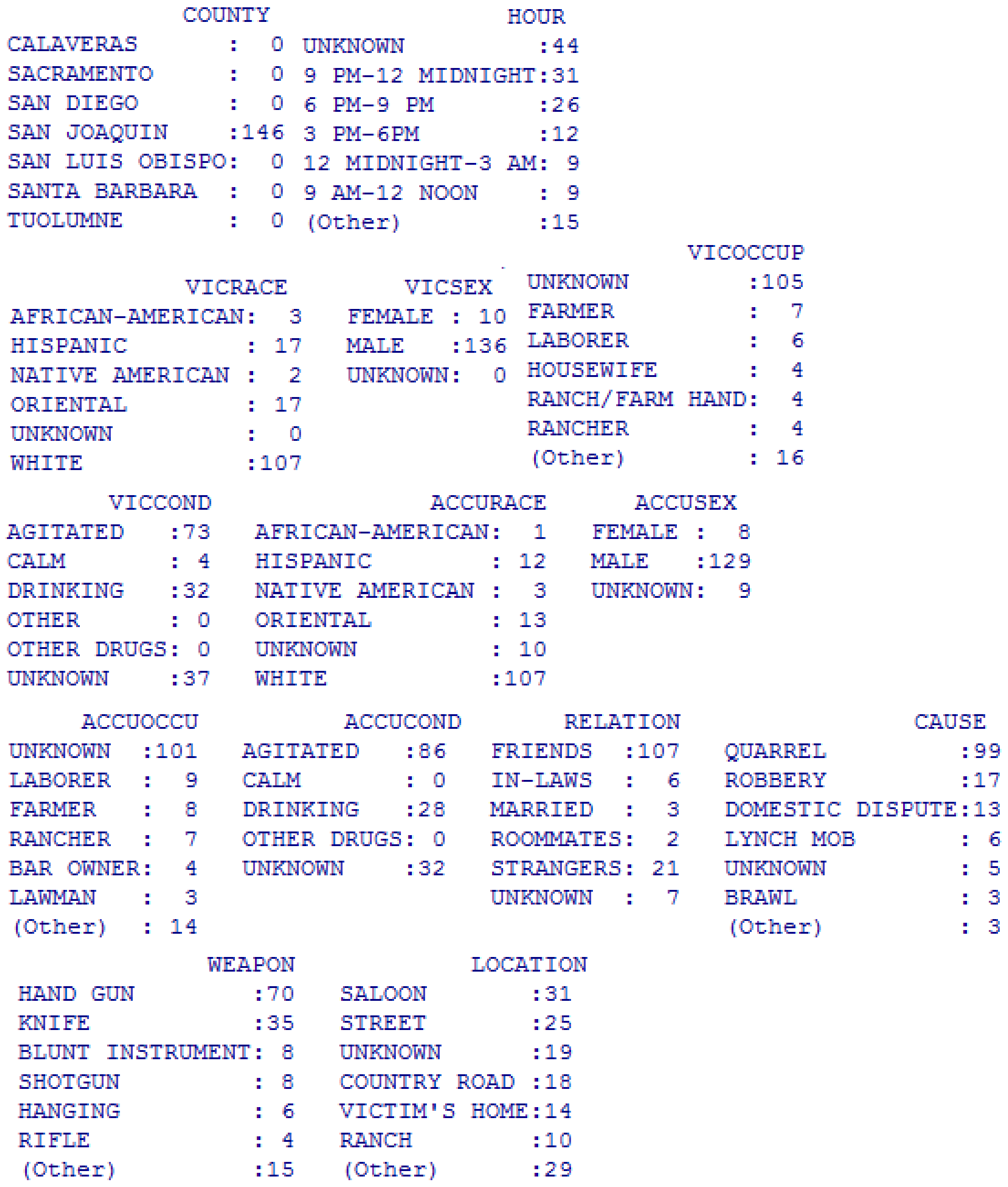


```
             COUNTY                 HOUR
CALAVERAS      :  0  UNKNOWN           :69
SACRAMENTO     :  8  9 PM-12 MIDNIGHT:55
SAN DIEGO      :220  6 PM-9 PM         :39
SAN JOAQUIN    :  0  9 AM-12 NOON      :19
SAN LUIS OBISPO:  2  3 PM-6PM          :16
SANTA BARBARA  :  1  12 NOON-3 PM      :15
TUOLUMNE       :  1  (Other)           :19
                                          VICOCCUP
            VICRACE        VICSEX  UNKNOWN   :78
AFRICAN-AMERICAN:  3   FEMALE : 24 LABORER   :22
HISPANIC        : 50   MALE   :208 RANCHER   :21
NATIVE AMERICAN : 38   UNKNOWN:  0 FARMER    :15
ORIENTAL        :  2               HOUSEWIFE:11
UNKNOWN         :  0               MERCHANT  :  9
WHITE           :139               (Other)   :76
       VICCOND              ACCURACE       ACCUSEX
AGITATED   :66   AFRICAN-AMERICAN:  7   FEMALE :  3
CALM       :28   HISPANIC        : 37   MALE   :211
DRINKING   :67   NATIVE AMERICAN : 51   UNKNOWN: 18
OTHER      : 0   ORIENTAL        :  2
OTHER DRUGS: 3   UNKNOWN         : 20
UNKNOWN    :68   WHITE           :115

    ACCUOCCU         ACCUCOND         RELATION                 CAUSE
UNKNOWN :99   AGITATED   :89   FRIENDS   :148   QUARREL           :135
LABORER :37   CALM       : 4   IN-LAWS   :  7   ROBBERY           : 41
RANCHER :20   DRINKING   :68   MARRIED   :  9   UNKNOWN           : 15
LAWMAN  :15   OTHER DRUGS: 3   ROOMMATES:  4   DOMESTIC DISPUTE: 14
FARMER  : 8   UNKNOWN    :68   STRANGERS: 47   KILLED BY POLICE: 12
MERCHANT: 7                    UNKNOWN   : 17   BRAWL             :  5
(Other) :46                                     (Other)           : 10
             WEAPON            LOCATION
HAND GUN          :86   COUNTRY ROAD :41
KNIFE             :39   RANCH        :36
BLUNT INSTRUMENT:23   VICTIM'S HOME:34
RIFLE             :22   SALOON       :25
SHOTGUN           :16   STREET       :25
UNKNOWN           :12   UNKNOWN      :11
(Other)           :34   (Other)      :60
```

From the fourth cluster we can highlight the following characteristics:

- It is a group represented by individuals of San Diego.
- The victim is mainly white and a man. The occupations that are most represented are the worker, farmers and ranchers.
- The relation between the victim and accused in most cases was that of friends.
- From within this group we would distinguish homicides from a rural population.

Cluster 5



```
                COUNTY                     HOUR
CALAVERAS        :  0 UNKNOWN           :44
SACRAMENTO       :  0 9 PM-12 MIDNIGHT:31
SAN DIEGO        :  0 6 PM-9 PM         :26
SAN JOAQUIN      :146 3 PM-6PM          :12
SAN LUIS OBISPO:  0 12 MIDNIGHT-3 AM: 9
SANTA BARBARA  :  0 9 AM-12 NOON      : 9
TUOLUMNE         :  0 (Other)           :15
                                               VICOCCUP
             VICRACE          VICSEX   UNKNOWN        :105
AFRICAN-AMERICAN:  3    FEMALE : 10  FARMER          :  7
HISPANIC           : 17    MALE    :136  LABORER         :  6
NATIVE AMERICAN :  2    UNKNOWN:   0  HOUSEWIFE       :  4
ORIENTAL           : 17                  RANCH/FARM HAND:  4
UNKNOWN            :  0                  RANCHER         :  4
WHITE              :107                  (Other)         : 16

       VICCOND                ACCURACE       ACCUSEX
AGITATED    :73   AFRICAN-AMERICAN:  1   FEMALE :  8
CALM         : 4   HISPANIC          : 12   MALE    :129
DRINKING    :32   NATIVE AMERICAN :  3   UNKNOWN:   9
OTHER        : 0   ORIENTAL          : 13
OTHER DRUGS: 0   UNKNOWN           : 10
UNKNOWN     :37   WHITE             :107

     ACCUOCCU          ACCUCOND         RELATION               CAUSE
UNKNOWN   :101   AGITATED    :86   FRIENDS   :107   QUARREL            :99
LABORER   :  9   CALM         : 0   IN-LAWS   :  6   ROBBERY            :17
FARMER    :  8   DRINKING    :28   MARRIED   :  3   DOMESTIC DISPUTE:13
RANCHER   :  7   OTHER DRUGS: 0   ROOMMATES:  2   LYNCH MOB         : 6
BAR OWNER:  4   UNKNOWN     :32   STRANGERS: 21   UNKNOWN           : 5
LAWMAN    :  3                     UNKNOWN   :  7   BRAWL             : 3
(Other)   : 14                                       (Other)           : 3

             WEAPON            LOCATION
HAND GUN          :70   SALOON        :31
KNIFE              :35   STREET        :25
BLUNT INSTRUMENT: 8   UNKNOWN       :19
SHOTGUN           : 8   COUNTRY ROAD :18
HANGING           : 6   VICTIM'S HOME:14
RIFLE              : 4   RANCH         :10
(Other)           :15   (Other)       :29
```

From the fifth cluster we can highlight the following characteristics:

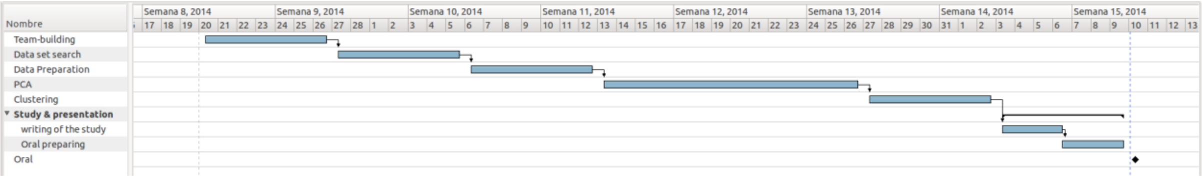
- It is a group represented by individuals of San Joaquin.
- The victim and the accused, in most cases are men and white. The majority occupation is farmer and worker.
- The relation between the victim and the accused was mostly friendship.
- The crime begins in a fight and ends with the victim at the bar or street.

10. Conclusions

- We observe that counties with higher population density and greater racial variety are opposed to counties with a higher level of education and less racial variety (Mainly white).
- The sex of the victim and the accused has significant relevance, particularly in cases of domestic violence.
- For both victims and accused their occupation is strongly related to their level of education and income. Crimes such as theft, fighting or killing are more strongly associated with poor education and low income, while other crimes such as those related to mafia are related to higher levels of study.
- We can identify that in the counties of San Luis Obispo, Santa Barbara, Calaveras and Tuolumne, the main profession of the victims and the accused is Miner. Something closely related to one of the activities in the area thing.
- In San Diego the disputes between farmers and/or ranchers are significant and constitute a large portion of the crimes.
- In populations with higher rate of non-white races and lower educational level there are more cases of domestic violence. And the victims are often in their own homes or in the street.

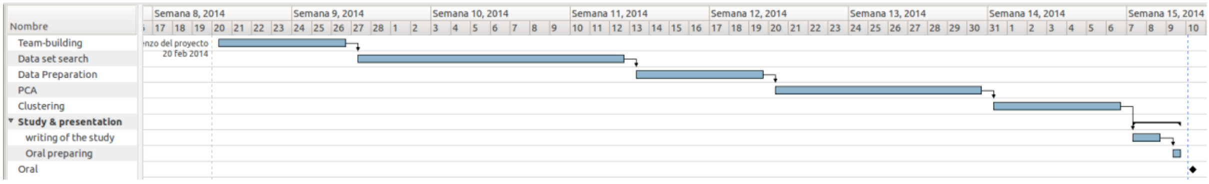
11. Planning of the work

At the beginning of the project we planned the days with the different tasks we had to perform. These tasks went from the consolidation of the group until the preparation of the oral defense of the study, through the stages of preparation of the data, PCA, clustering etc. Not to mention the writing of this document. At the end, the planning of the project was reflected in this Gantt diagram.



The division of tasks during the performance of this study done among us, was made so that all team members actively engaged in all phases of project, so that we are all assigned to the same task for as long as this lasts.

On the other side, once finished the study, we see that the planned schedule has not been exactly the actual calendar, which has been reflected by the following Gantt diagram.



The main differences and deviations from the original planning were mainly caused by the delay of a week to select a data set that would allow us to do the study. With this delay we were forced to reduce the time spent on PCA in four days. And finally the time planned for the drafting of the document and the preparation of the defense of the project have taken less time than expected.