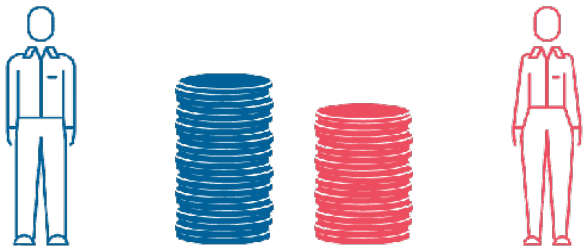


Gender Salary Gap

Data mining Project 1

Members

Joan Areal Azuara

Laia Bonilla Pérez

Miguel Gutiérrez Jariod

Fujie Mei

Alessandro Scannavini

0. Motivation, general description and analysis

The existence of the gender pay gap is a topic of much debate and research. Some studies suggest that the gender pay gap is a myth or exaggerated, while others argue that it is a persistent and systemic problem that affects women's economic security and well-being.

Understanding the nature and extent of the gender pay gap, and identifying the factors that contribute to it, is an important goal for policymakers, employers, and society as a whole. By addressing the gender pay gap, we can improve women's economic security, promote gender equality, and support the overall economic growth and productivity.

Our project aims to investigate the gender pay gap using a dataset obtained from Kaggle. The dataset contains information about the earnings of men and women in that industry, as well as other relevant variables, such as work experience, education and job title.

Our main objective is to analyze the data and determine whether there is a significant gender pay gap and analyze the factors that may affect earnings.
For this purpose we will use data mining techniques such as PCA (Principal Component Analysis) and clustering:

- PCA is a statistical technique used for identifying the most meaningfull features of the dataset, allowing a dimensionality reduction of the dataset that can simplify the analysis and highlight the key factors that contribute to the gender pay gap
- Clustering is a statistical technique that groups individuals or data points into classes (i.e. clusters) based on their similarities. By clustering the data, we can identify patterns that may be relevant to understanding the gender pay gap. For example, we may identify clusters of workers who have similar earnings, work experience, education, or other characteristics, and investigate whether there are gender differences within or across these clusters.

1. Metadata

1.1 Numerical

Variable	Type	Range	Units	Missing (%)	Meaning	Role
numprec	numerical	[1, 16]	Persons	0	Number of person records following	Explanatory
age	numerical	[25, 64]	Years	0	Age of the individual	Explanatory
annhrs	numerical	[10, 5148]	Hours	0	Annual hours worked	Explanatory
hrwage	numerical	[2.3, 591]	Dollars	0	Hourly wage	Explanatory
incwage	numerical	[38, 1099999]	Dollars	0	Wage and salary income	Response
uhrswork	numerical	[1, 99]	Hours	0	Usual weekly hours worked	Explanatory
wkswork1	numerical	[1, 52]	Weeks	0	Weeks worked last year	Explanatory

1.2 Categorical

Variable	Type	Modalities	Missing (%)	Meaning	Role

region	categorical	New England Division, Middle Atlantic Division, East North Central Div., West North Central Div., South Atlantic Division, East South Central Div., West South Central Div., Mountain Division, Pacific Division	0	Living division of the individual	Explanatory
metro	categorical	Not in metro area, In metro area, central / principal city, In metro area, outside central / principal city, Central / Principal city status unknown, Not available	0,76	Metropolitan central city status	Explanatory
relate	categorical	Head/ Householder Spouse, Child, Parent, Sibling, Grandchild, Other Relatives, Unmarried Partner, Housemate/ Roomate, Roomers/ boarders/lodgers, Other non-relatives	0	Relationship of the individual to the household head	Explanatory
sex	binary	Male, female	0	Sex of the individual	Explanatory

race	categorical	White, Black/African American, American Indian or Alaska Native, Chinese	0	Race of the individual	Explanatory
marst	categorical	Married spouse present, Married spouse absent, Separated, Divorced, Widowed, Never married/ single	0	Maritial status	Explanatory
bpl	categorical	Guam, Puerto Rico, Canada, Mexico, Costa Rica, … (108 modalities)	0,06	Birthplace	Explanatory
sch	categorical	None, Grade 2, Grade 5, Grade 7, Grade 9, Grade 10, Grade 11, Grade 12, Some collage, Associate's degree, Bachelor's degree, Advanced degree	0	Educational attainment recode	Explanatory

educ99	categorical	No school, 1st-4th grades, 5th-8th grades, 9th grade, 10th grade, 11th grade, 12th grade, High school, Some collage, Associate's degree 1, Associate's degree 2, Bachelor's degree, Master's degree, Professional degree, Doctoral degree	0	Educational attainment	Explanatory
empstat	binary	At work, Has job not at work now	0	Employment status	Explanatory
classwkr	categorical	Wage/salary, private sector, Federal govt employee, State govt employee, Local govt employee	0	Class worker	Explanatory
ftype	categorical	Primary family, Nonfamily householder, Related subfamily, Secondarted subfamily, Unrelay individual	0	Family type	Explanatory
adj_ind	categorical	Crop production, Animal production, Forestry except logging, Logging … (219 modalities)	0	Industry of the individual	Explanatory

adj_occ2name	categorical	Managers, Production, Admin Support, Construction, Extraction, Installation, Computer and Math Technicians, Community and Social Workers, Business Operators, Transportation and materials moving, Provide Service adj_occupations, Office and Admin Support, Farming Phishing and Forestry, Lawyers, Judges, Physicians and dentists	41,56	Occupation of the individual	Explanatory

2. Preprocessing

In this part of the project, we have processed the origin dataset. So we have done some transformations to accomplish the rules given. Here we show which are these and how we have proceeded.

Firstly, we analyzed our data checking that was from dataset class. Having that in order, our original dataset was composed by 344287 rows and 243 columns, so we had to remove some of that to accomplish the 5000 maximum rows and 30 maximum columns.

Also, there was a lot of data among different years, so we decided to focus only on 2013 participants. Moreover, to have a heterogeneous set between men and women, half of the dataset is composed per one sex and the other half per the other sex, so we can extract information about the salary gap considering both parts in a similar amount of people.

Finally, our dataset was composed by 5000 rows and 21 columns. About this last one, we had to look for the main interesting and also considering accomplishing the 7 numerical, 7 categorical and 2 binary at least. So we took them by knowing what does the variable measures.

Once we have our dataset prepared, before treating it properly, we had to consider the amount of nullable data that was included in it. For each variable, we count how many NA's data had. Also, we show the total of the nullable in the data and the percentage that it corresponds.

Originally, we had a variable that had an 81% of nullable values, so we decided to remove it. Moreover, there weren't too many nullable values, the main one to consider was adj_occ2name, which was 41%. Only categorical variables had NA values.

80
60
40
20
0
Afghanistan
Bangladesh
Caribbean
Dominica
Fiji
Guinea
Italy
Lithuania
Pakistan
Senegal
Syria
ited States

20
15
10
5
0
s Operators
al Workers
echnicians
Installation
& Forestry
l Operators
nd dentists
Managers
hin Support
Production
ccupations
als moving

30
20
10
0
s unknown
rincipal city
rincipal city
metro area

Once we had the knowledge about the number of nullable in the different variables, we proceed to change the NA for the main value obtained when we apply KNN. We verify it with plots to see that the structure of the proportion graphs stays the same before and after the imputations.

The number of neighbors chosen is K=71, which is $\sqrt{5000}$, 5000 being the number of individuals we have, rounded to the nearest odd number. We found this criteria in this article: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/, which states the following: "Choice of k is very critical – A small value of k means that noise will have a higher influence on the result. A large value make it computationally expensive and kinda defeats the basic philosophy behind KNN (that points that are near might have similar densities or classes) .A simple approach to select k is set $k = \sqrt{n}$ "

Finally, we had run the detection of nullable values just to confirm that now all the variables have a corresponded value.

Factorial for the categorical values:

Now that we have completed the principal modifications and the data is ready to have some treatment, we have focused on the categorical variables to make understandable all the plots that had been produced in the descriptive analysis part. To realize it, we had to take knowledge of the different categories that each variable can take, so once they were analyzed we can proceed to do the corresponding factoring. Here we don't show all the code because it is quite long, as there are 14 categorical variables and some of them can adopt many categories, but the main procedure it's the one explained above.

3. Descriptive analysis

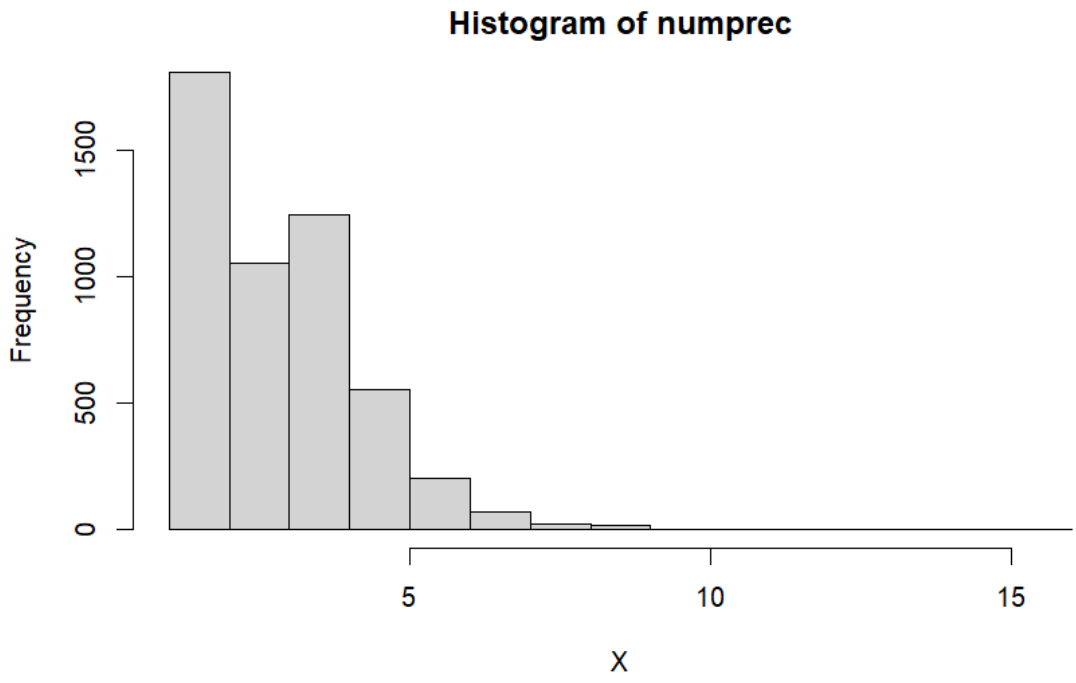
In this section of the document, we will perform a descriptive analysis of the variables of our dataset.

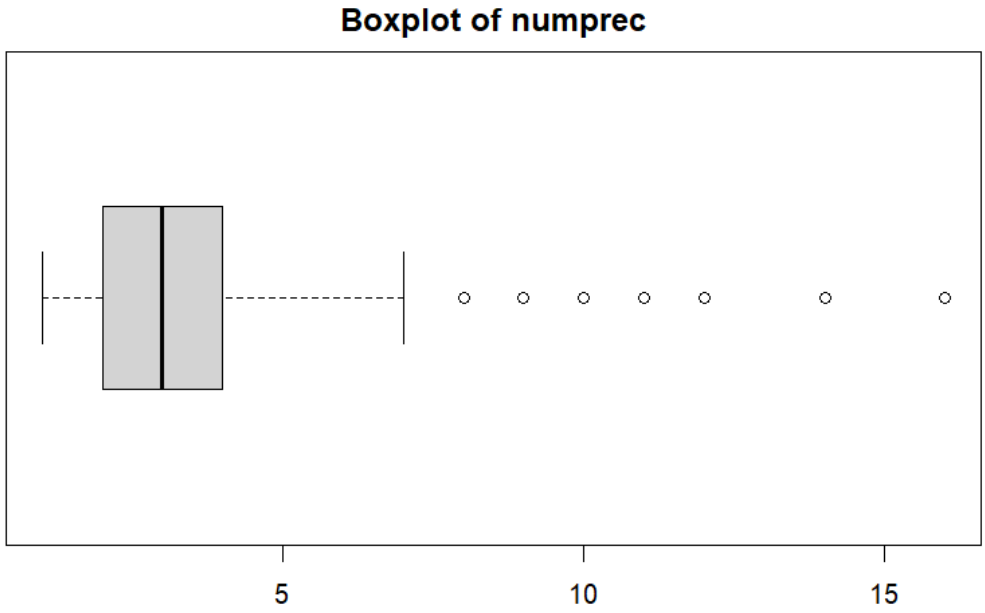
We will start by analyzing the numerical ones, with a histogram, a boxplot, a table with a statistical summary and a brief description.

For analyzing the categorical ones, we will include a table with the frequency of every value a variable can take, a pie chart, and a bar plot.

3.1 Numerical values

3.1.1 Numprec



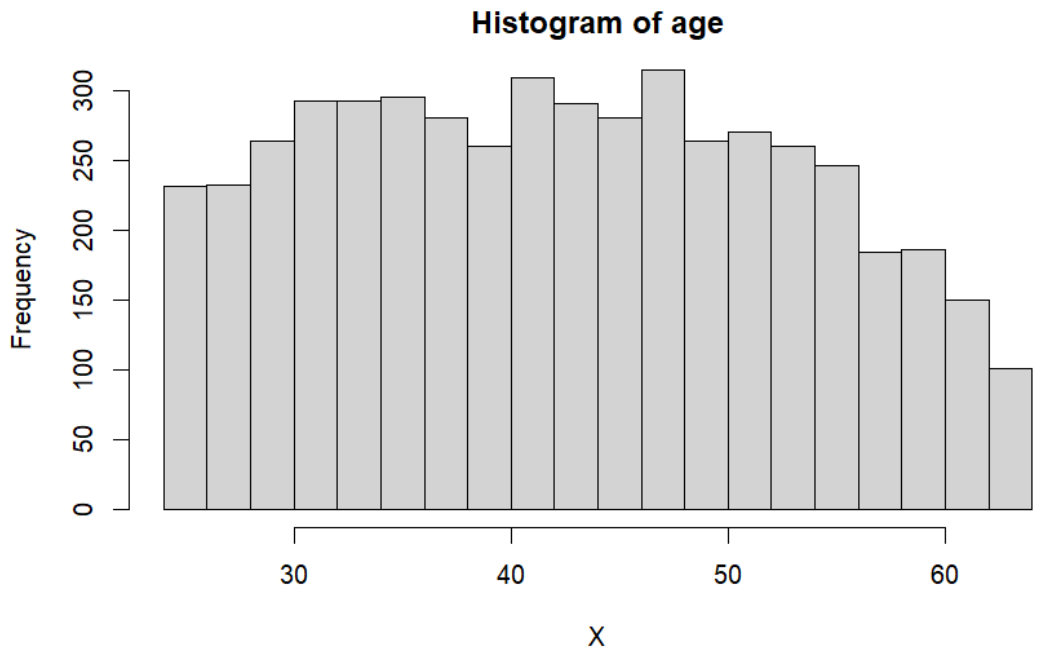


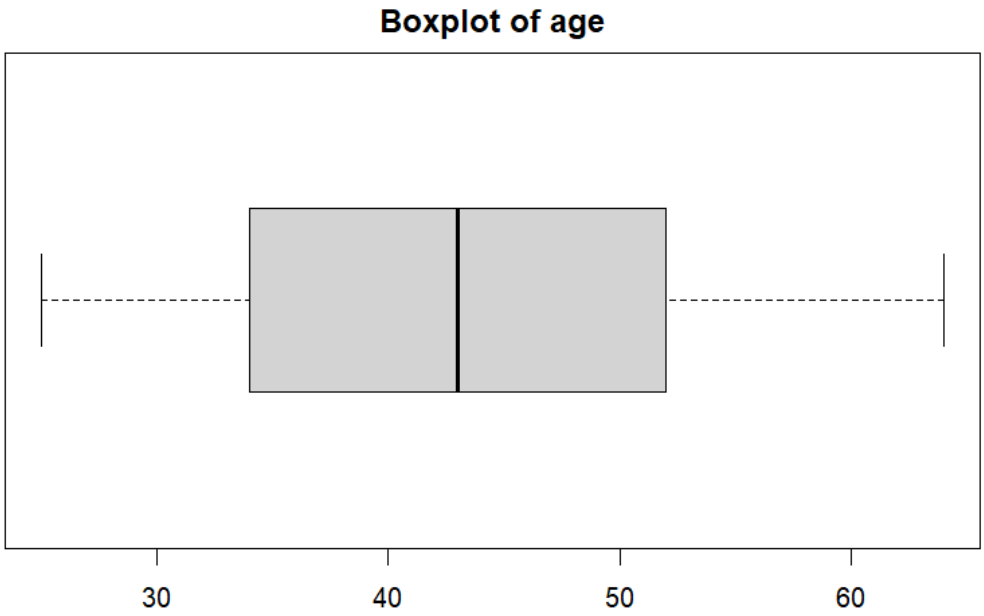
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
1.00	2.00	3.00	3.25	4.00	16.00	1.52	0.46

Description:

As we can see, observing the values for the 1st and 3rd Quantiles, this distribution is symmetrical. The boxplot also allows us to see that there are some outliers. There are some outliers, but compared to other numerical variables we have studied, it's a small amount.

3.1.2 Age



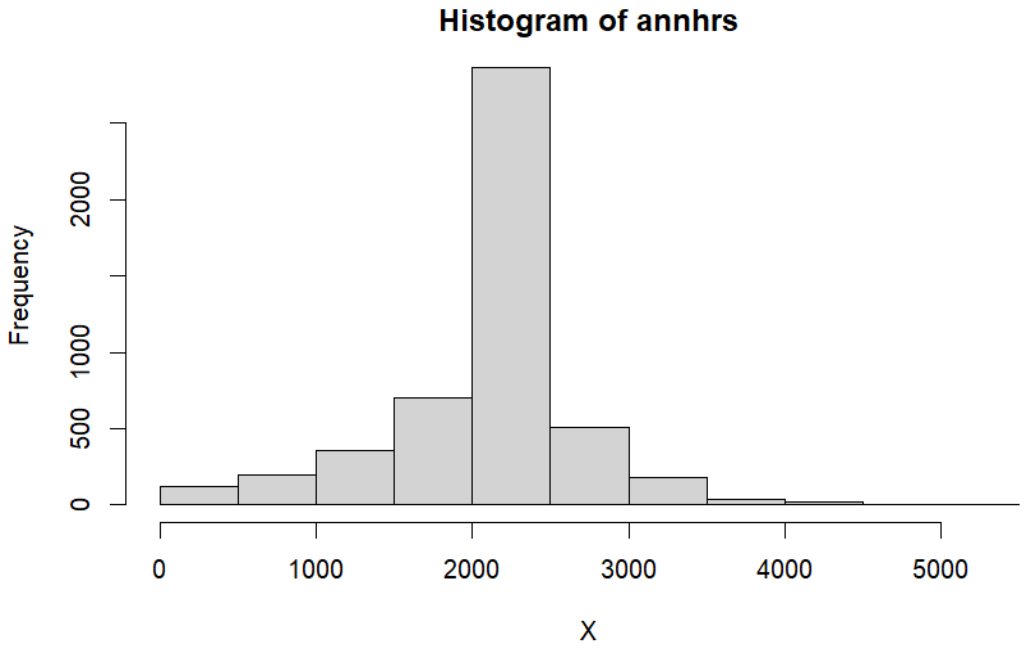


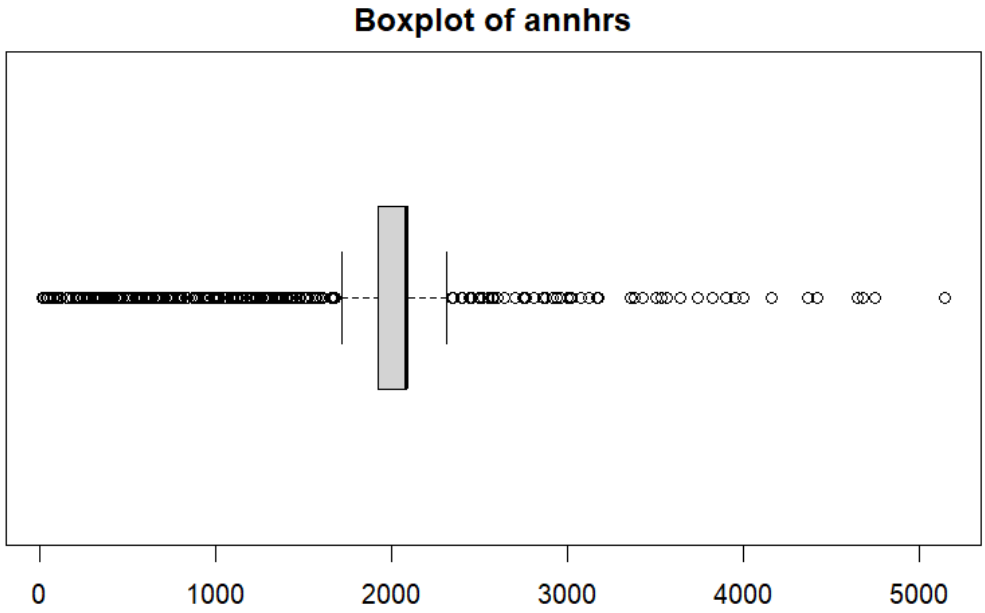
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
25.00	34.00	43.00	43.04	52.00	64.00	10.56	0.24

Description:

As we can see observing the values for the 1st and 3rd Quantiles, this distribution is symmetrical.
The histogram allows us to see that most age ranges are quite evenly represented in our dataset, with the exception of the ones over 60, which have fewer individuals.
In this case we don't have any outliers.

3.1.3 Annhrs



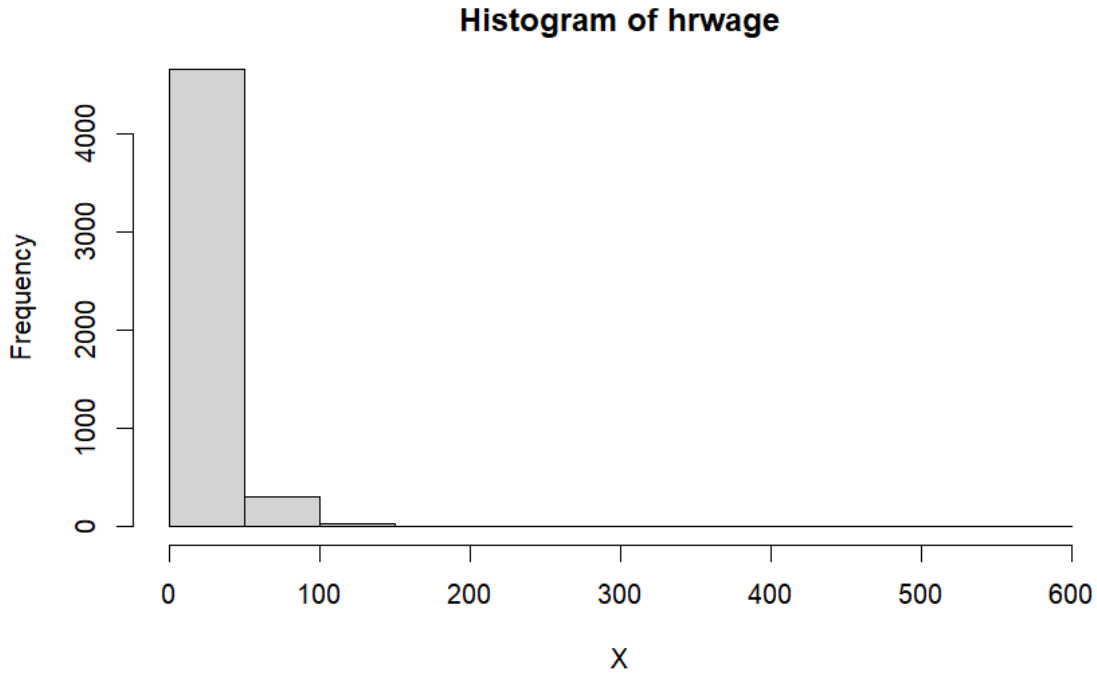


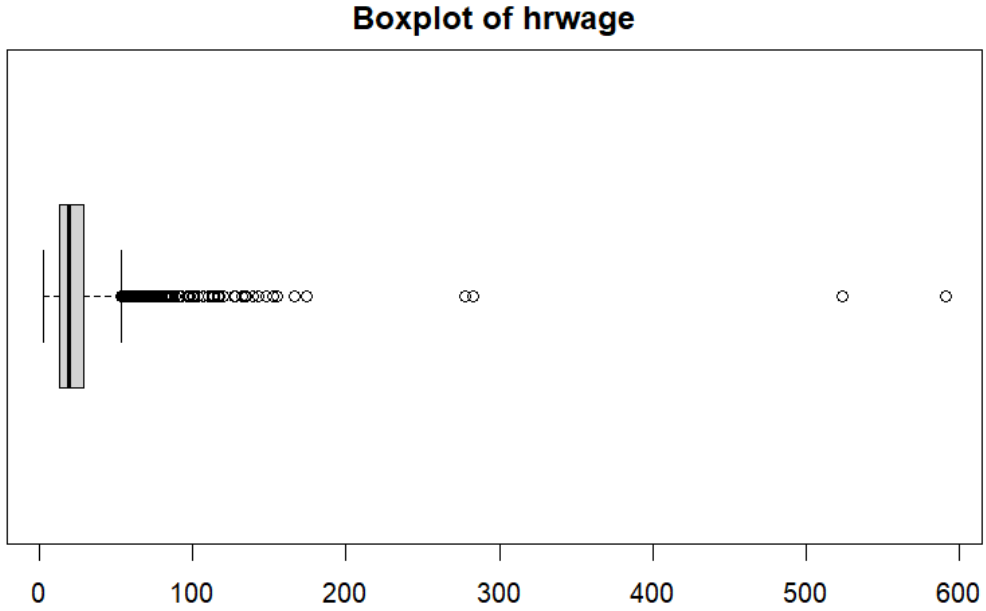
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
10	1924	2080	2023	2080	5148	592.80	0.29

Description:

The histogram of annual worked hours allows us to see that more than half of the individuals are in the range of 2000 to 2500 worked hours, the rest are quite dispersed but most of them are still close to that range.
This variable has a big amount of outliers, and a great deviation in the values. In this case, there are outliers on both sides of the median.

3.1.4 Hrwage



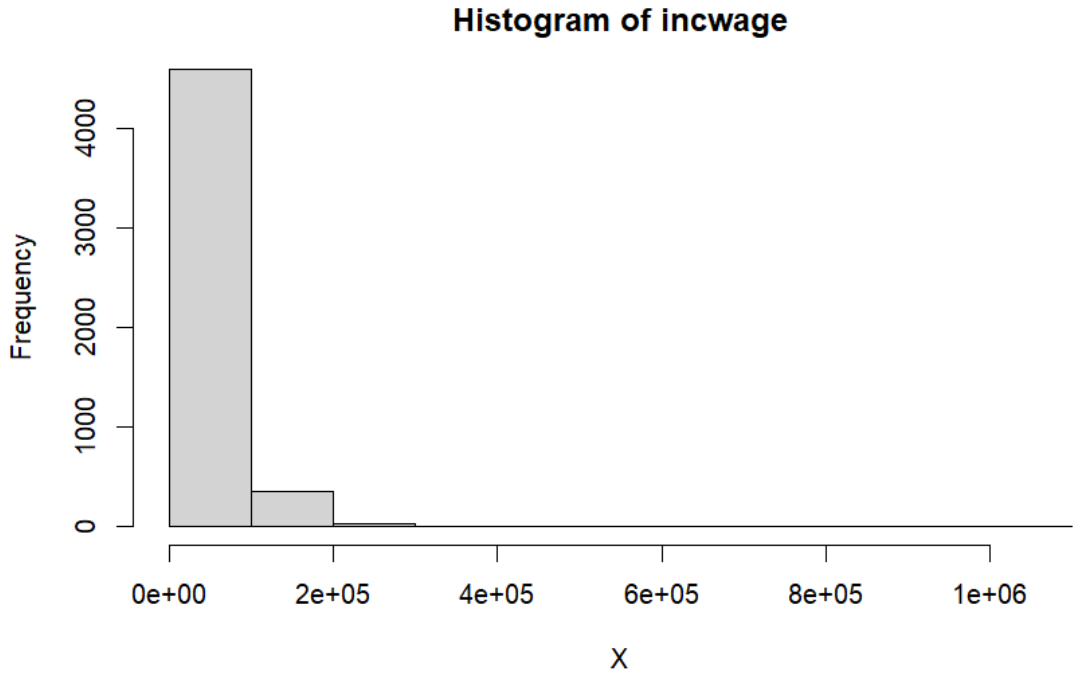


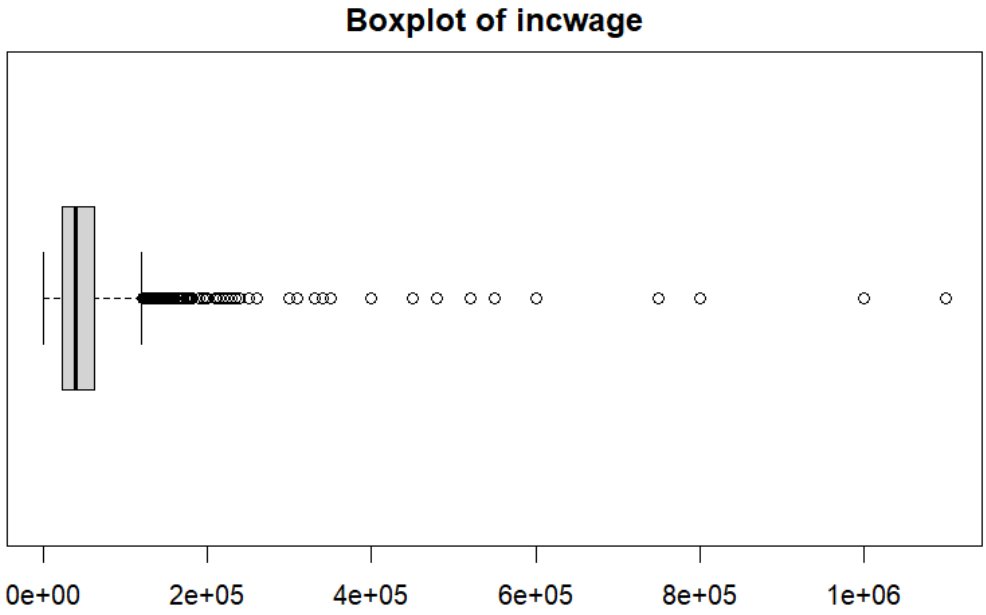
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
2.38	12.58	19.23	24.50	28.84	591.34	22.07	0.90

Description:

The boxplot shows that there is a considerable amount of outliers, there's a big amount of them that are not so deviated, but there are also a couple that are very far from the rest of values.
As we can see the histogram shows us that most of the individuals are in the range of 0 to 50 dollars per hour, the outliers make the scale of the histogram a bit unfitting to see the difference between the individuals inside that range, after all a difference of 10 to 49 dollars per hour is quite big.

3.1.5 Incwage



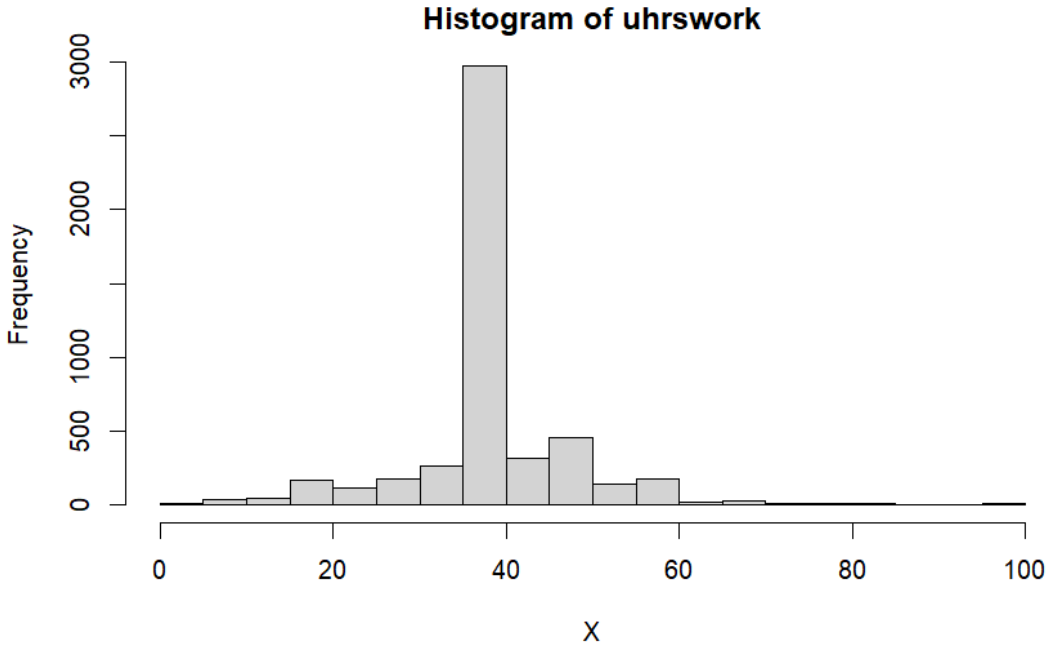


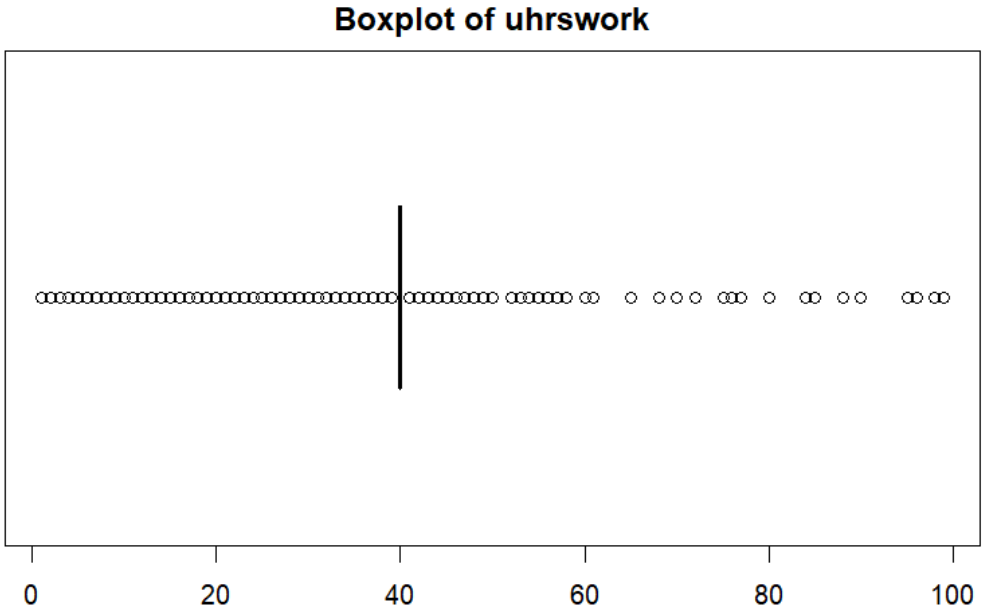
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
38	24000	39000	51862	63000	1099999	59862.51	1.15

Description:

This histogram has the same problem as the hourly wage one, the outliers mess up the scale and make it a bit difficult to really see how most of the individuals are distributed inside that range.

3.1.6 Uhrswrk



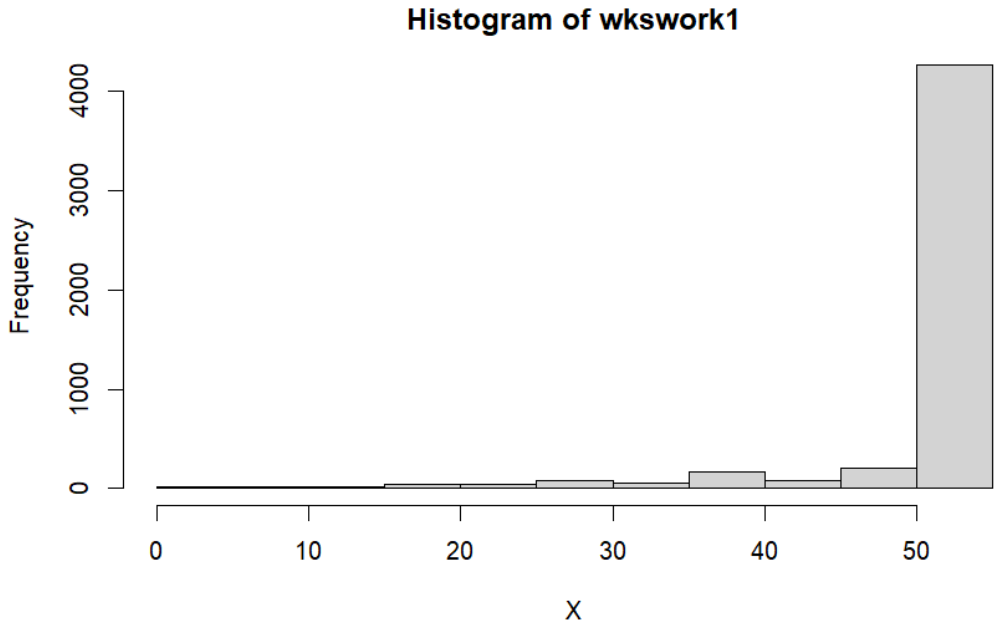


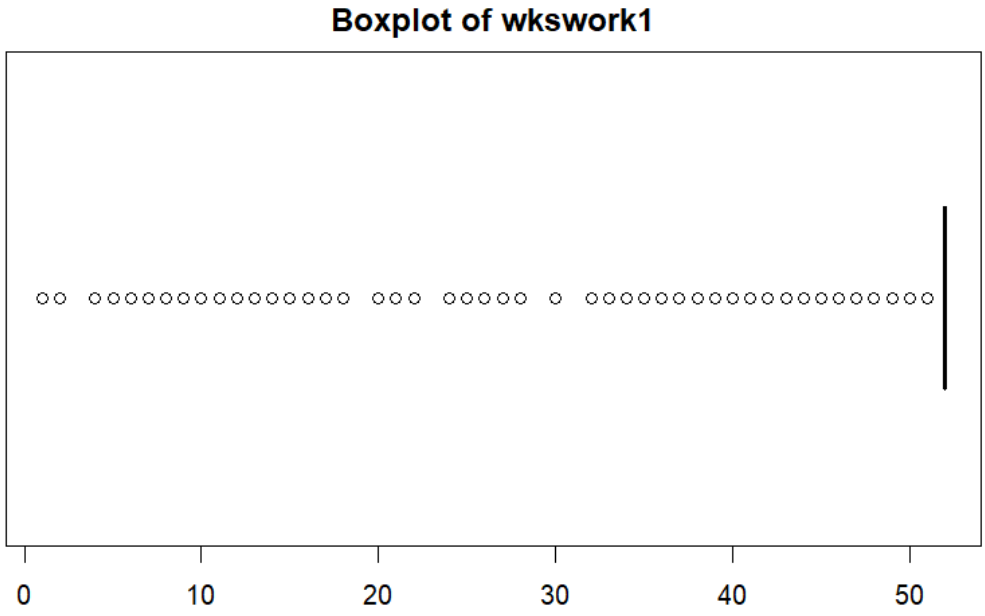
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
1.00	40	40	40.56	40.00	99.00	10.10	0.24

Description:

This histogram shows us the majority of the individuals on our dataset usually work 35-40 weeks per hour. The rest can be helpful to identify individuals that work part-time or overtime.
There are so many individuals that work the same amount of hours that the median and the 1st and 3rd Quartiles have the same values. There's also a considerable amount of outliers to both sides of it.

3.1.7 Wkswork1





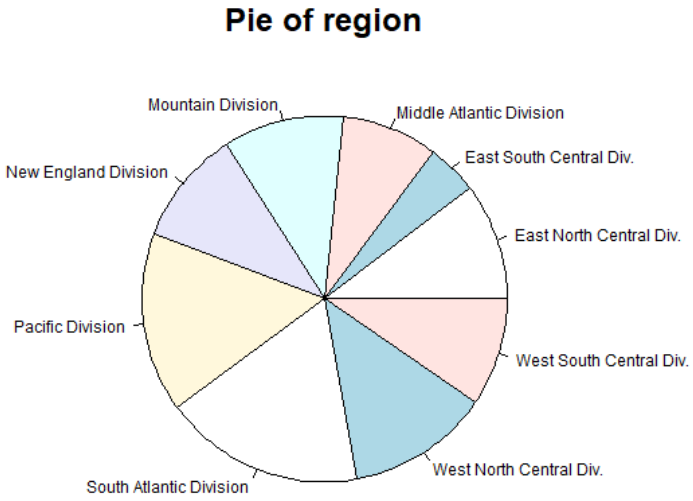
Min	1st Qu	Median	Mean	3rd Qu	Max	SD	VC
1.00	52.00	52.00	49.54	52.00	52.00	7.59	0.15

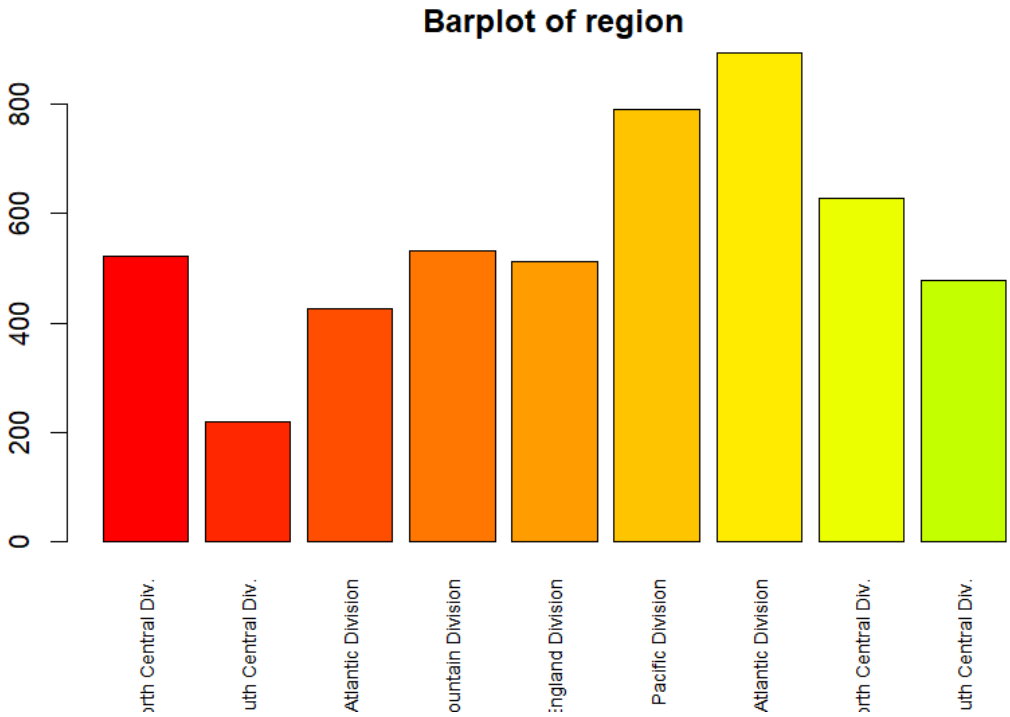
Description:

Most of the individuals on our dataset work between 50 and 53 weeks per year, given the year has around 53 weeks, we can conclude that most of them work the whole year.
For this variable the mean is equal to the 1st and 3rd Quantiles.
There are outliers, all of them with a lower value than the median.

3.2 Categorical variables

3.2.1 Region

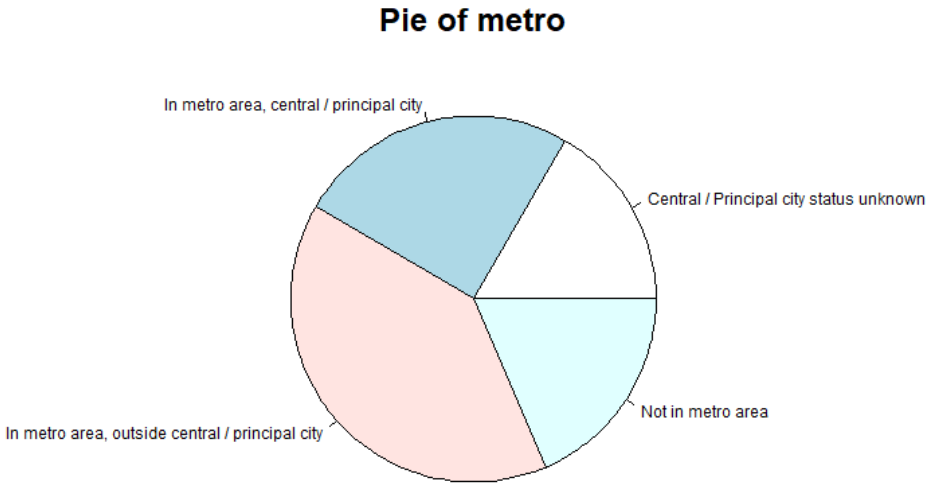


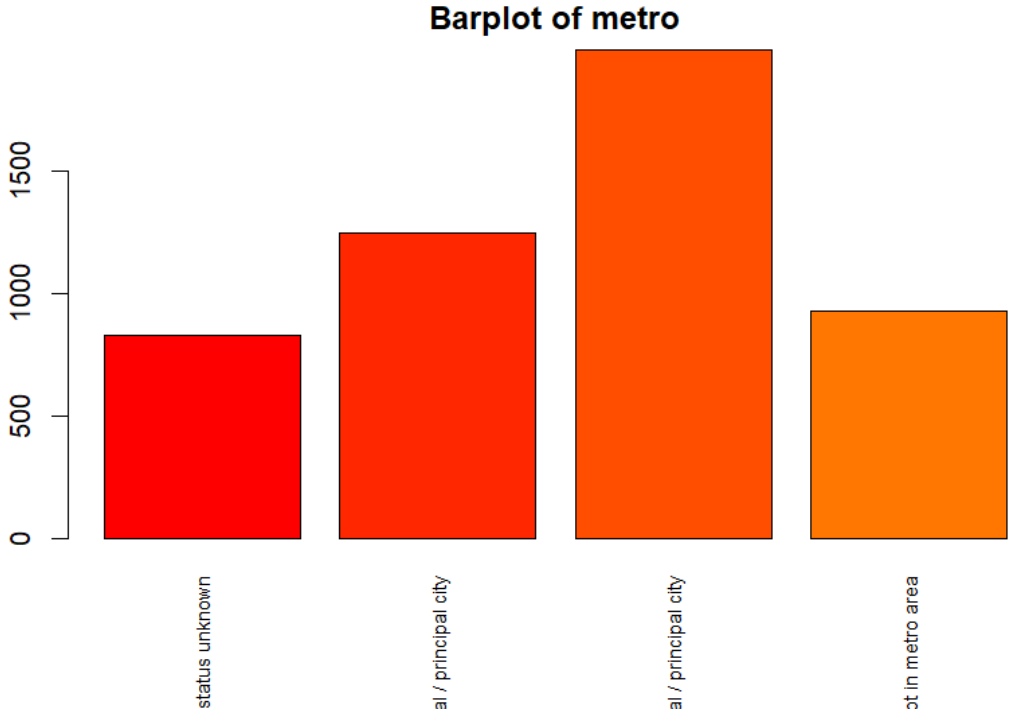


Description:

These graphs allow us to see the different regions the individuals belong to and how they are distributed. They are distributed according to which of the permanent divisions of the army organization they live in. We can see there are two predominant ones, and one with a lower number of individuals than the rest.

3.2.2 Metro

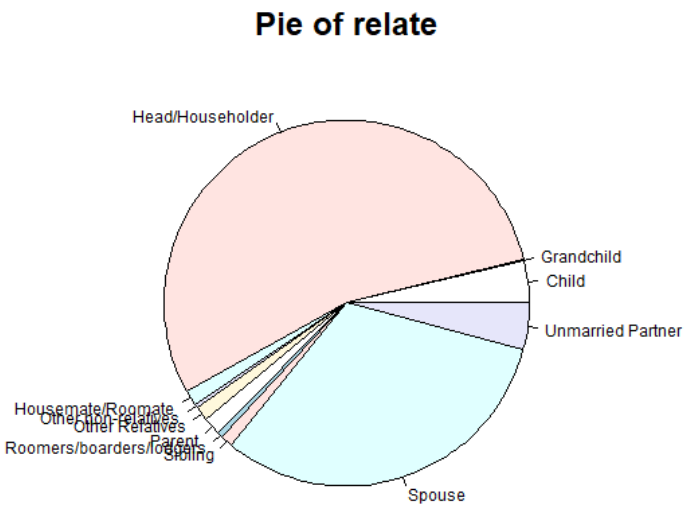


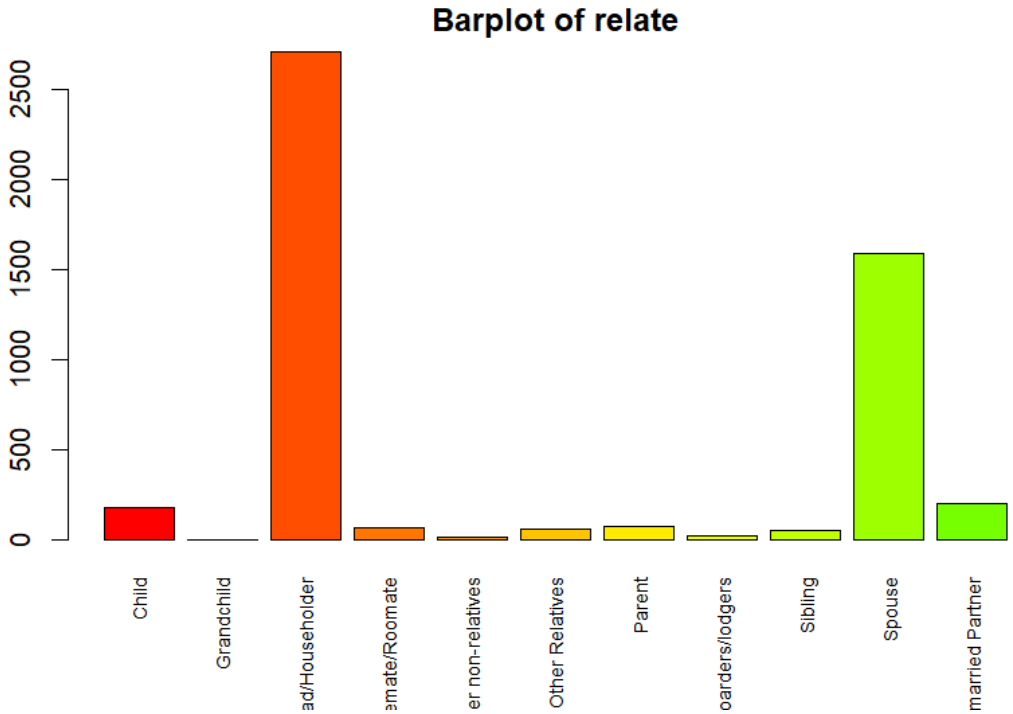


Description:

These plots allow us to see if the individuals live in metropolitan areas or not, and also what kind of metropolitan area they live in. As we can see the majority of individuals live in metropolitan areas, with more of them living outside the central city than inside. However, there is a considerable amount of people who live in a central city but with principal city status unknown, if we knew those, things could change.

3.2.3 Relate


Pie of relate
Head/Householder
Grandchild
Child
Unmarried Partner
Housemate/Roomate
Other non-relatives
Other Relatives
Parent
Roomers/boarders/foster
Sibling
Spouse



Description:

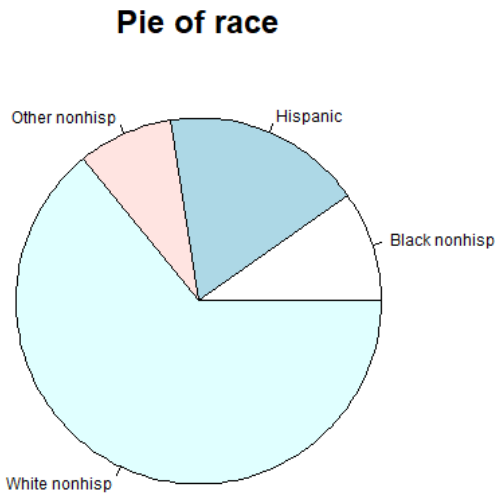
These plots allow us to see the relationship of the individuals with the head of their household. As we can clearly see, the main two groups of individuals are the ones that are the head of the household themselves, and the ones that are married to the head of the household, with the first ones representing more than half of the individuals in our dataset.

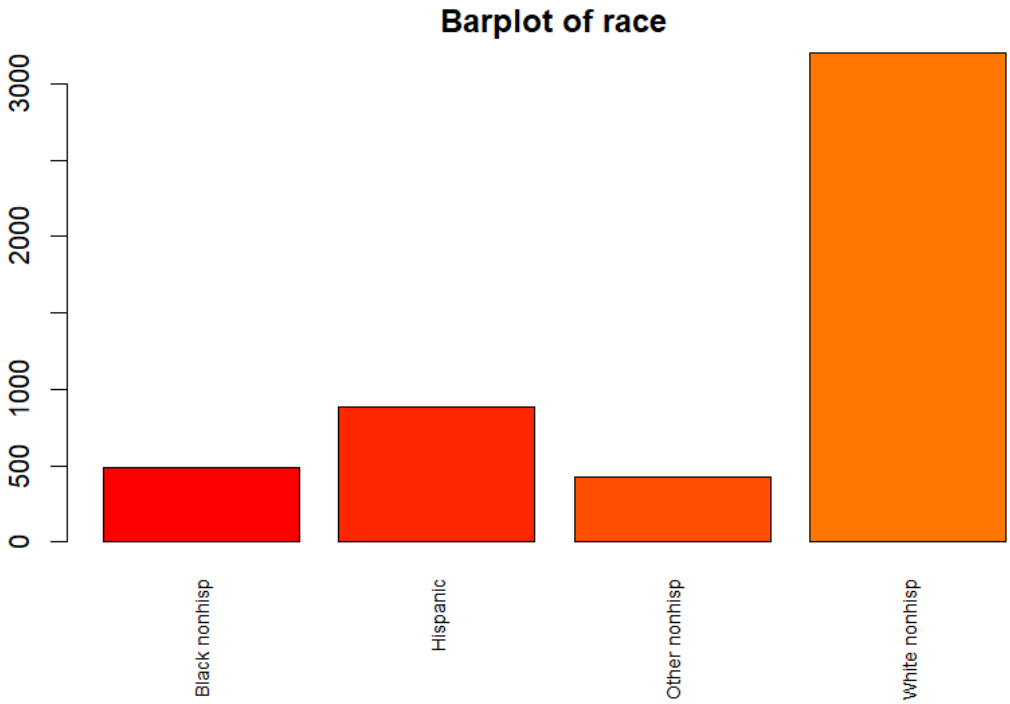
3.2.4 Sex

Description:

These plots show the sex of the individuals in our dataset. It's not a coincidence that they are 50-50, we picked our individuals out of a bigger dataset and we intentionally made it, so there would be an equal number of each, given the thematic.

3.2.5 Race

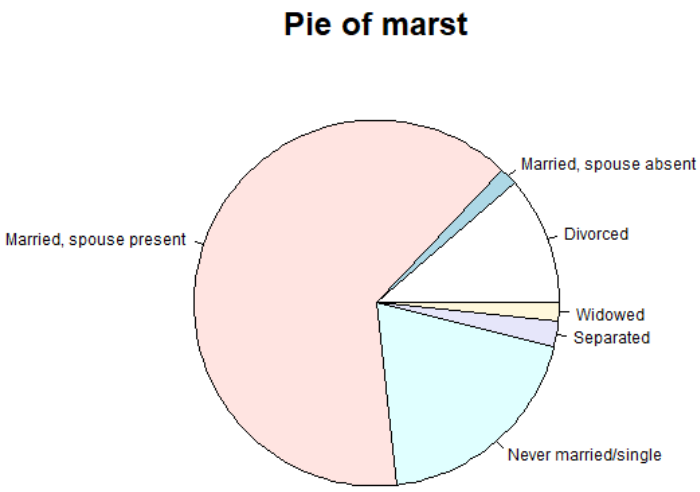


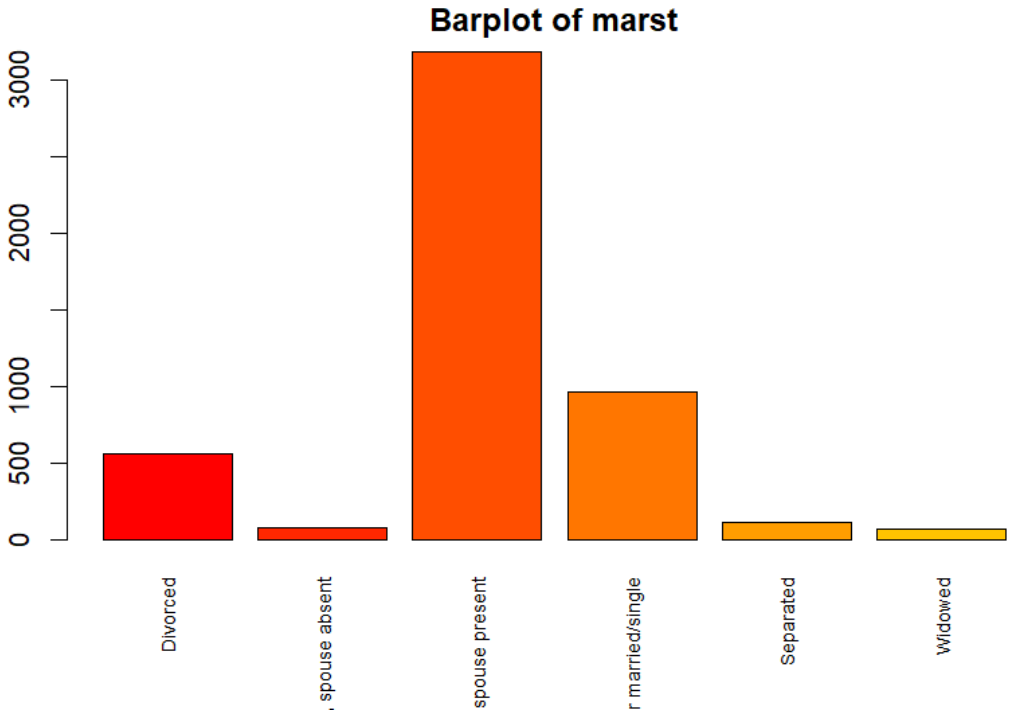


Description:

The majority of the individuals in our dataset are white. Although that is the most represented group, the others still have a considerable amount of individuals.

3.2.6 Marst

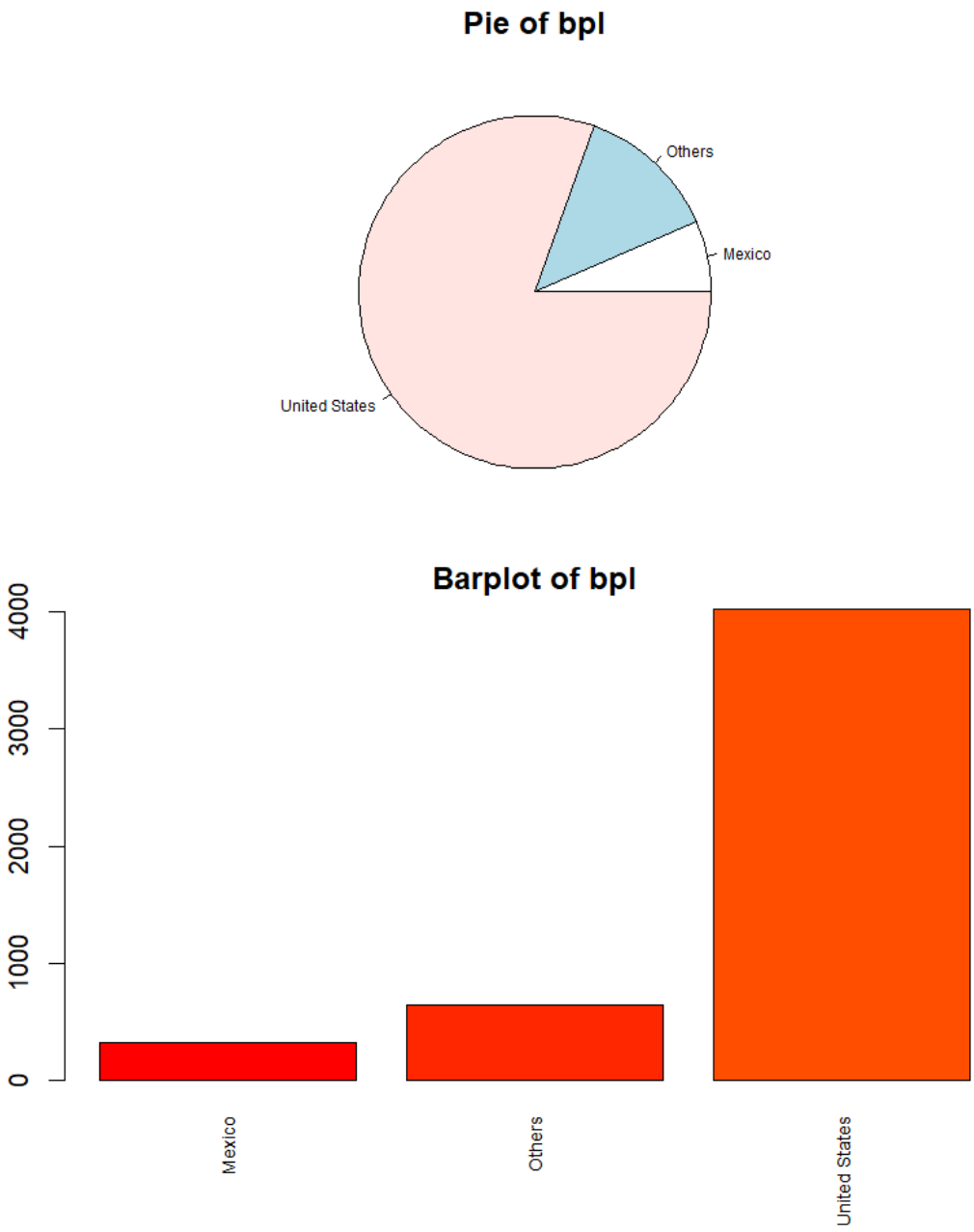




Description:

The graphs allow us to see that the majority of the individuals are currently married, even if there is a significant percentage of “Never married/single” and divorced.

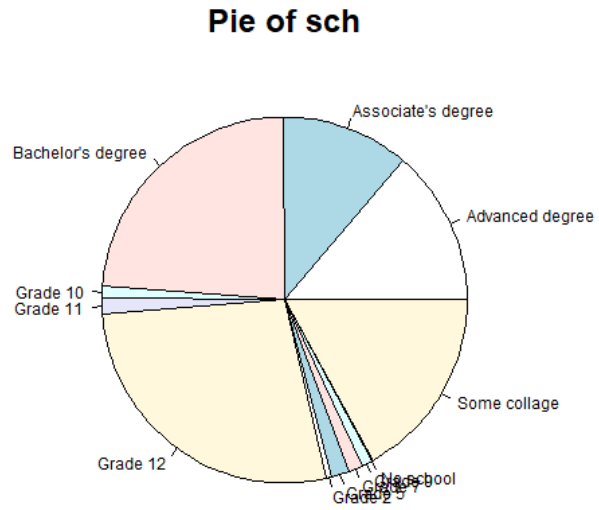
3.2.7 Bpl

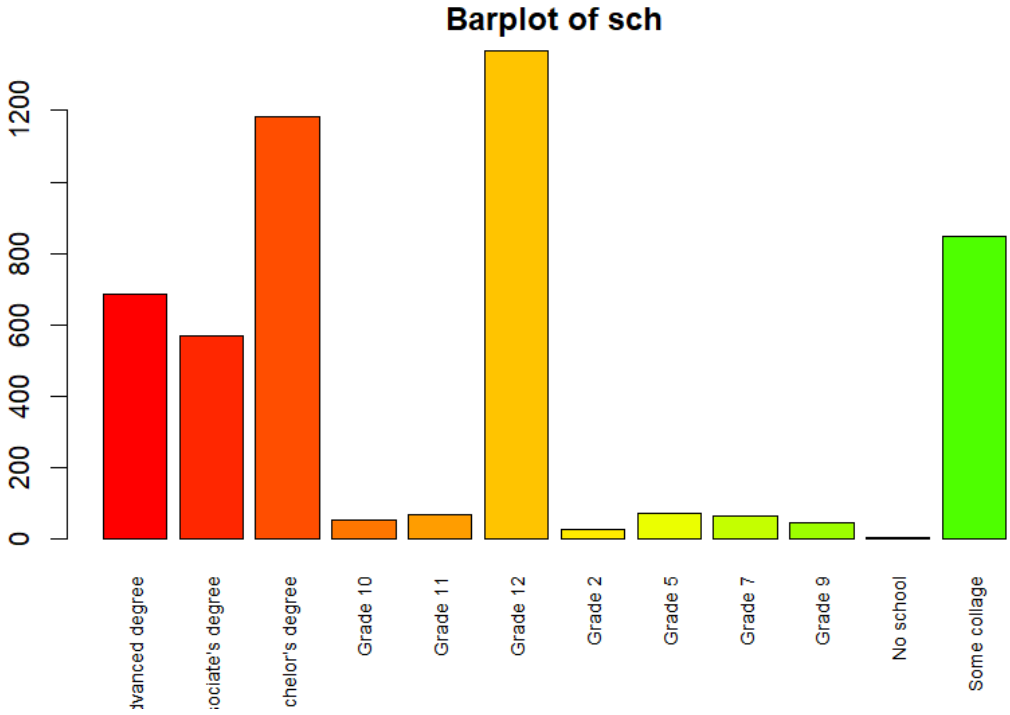


Description:

The graphs allow us to see that the majority of the individuals are from the United States, even if there is still a significant percentage of individuals from Mexico. All the other birthplaces could be grouped into an “Others” label for reaching a significant percentage.

3.2.8 Sch


Pie of sch
Associate's degree
Bachelor's degree
Advanced degree
Grade 10
Grade 11
Some collage
Grade 12


Barplot of sch
0
200
400
600
800
1200
Grade 10
Grade 11
Grade 12
Grade 2
Grade 5
Grade 7
Grade 9
No school
Some collage

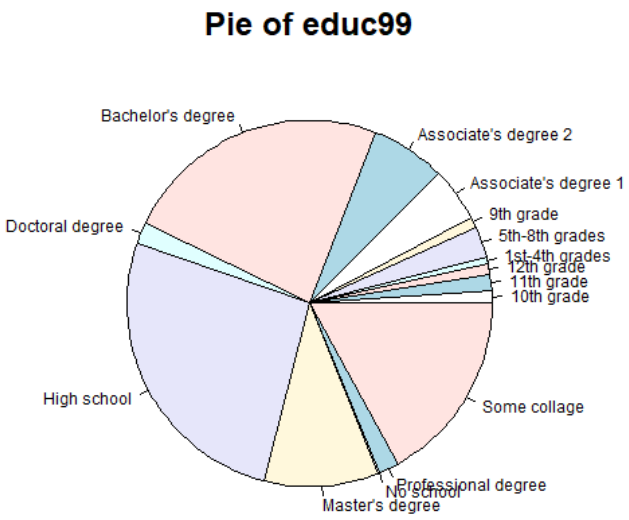
Description:

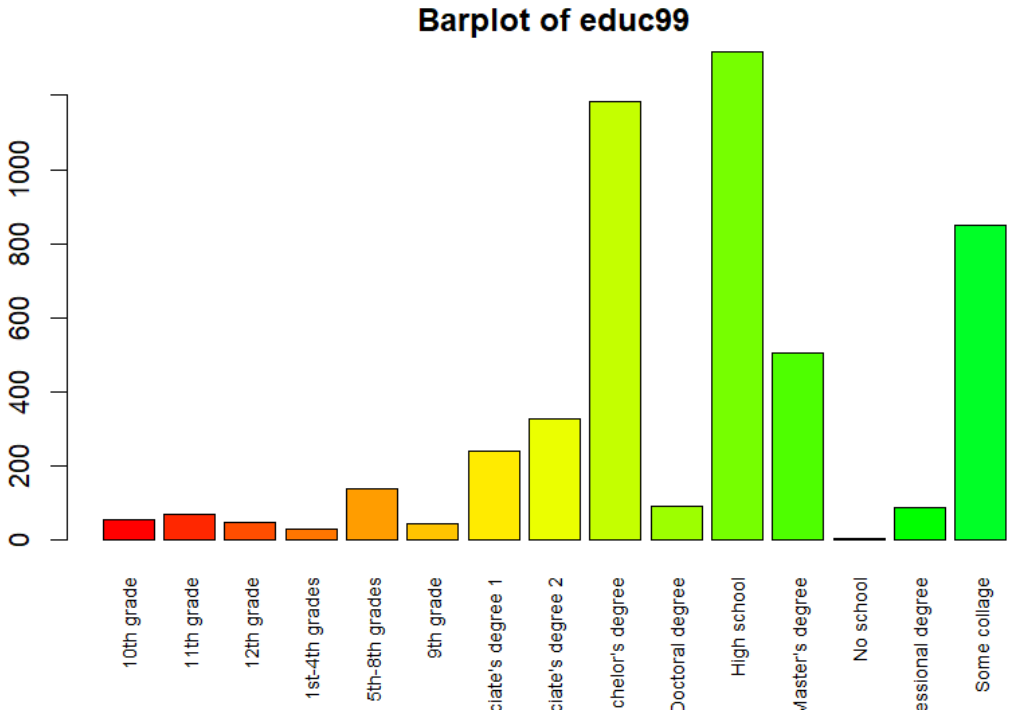
The pie and barplot allow us to see that around 50% of the individual's level of education is a Bachelor's, Associate's, or Advanced degree. There is also a considerable amount of individuals that attended some college, All these people pursued further education after highschool.

Around 25% of the individuals finished their education after Grade 12, which means that they didn't pursue some further education after highschool. The amount of individuals that finished their education before that is pretty small, but not so much if we grouped them all together. Still the group of people who didn't at least graduate high school is quite small.

If we are looking at it individually, without grouping it further, people who finished their education after Grade 12 represent the group with the biggest amount of individuals.

3.2.9 Educ99

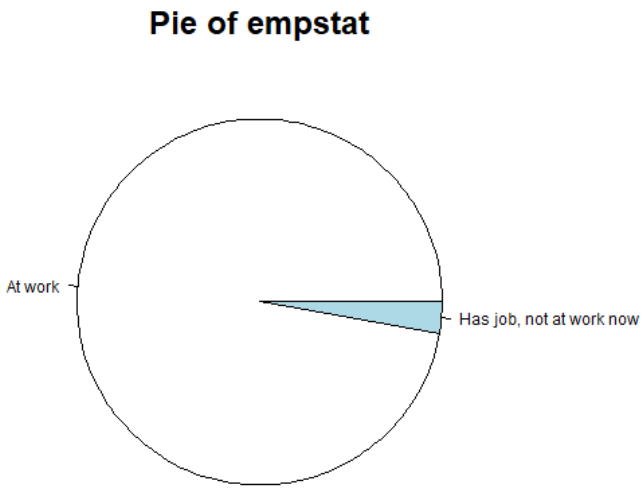


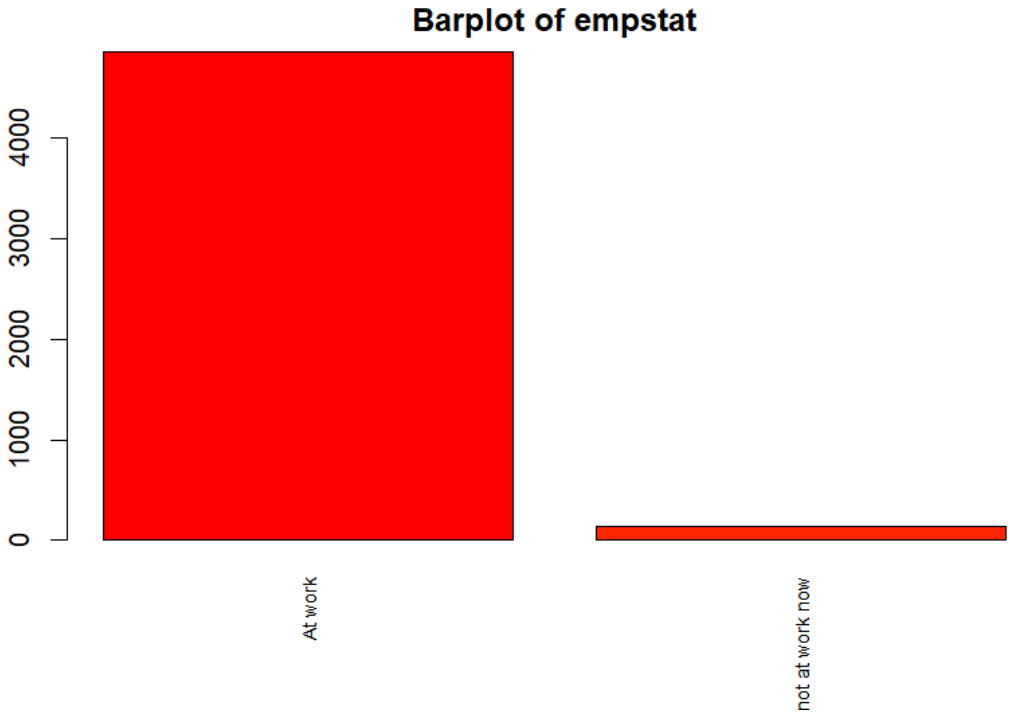


Description:

The graphs allow us to see that the majority of the individuals have a high educational level, because more than three fourth of the individuals have at least a high school degree, while more than a half of them have a university degree.

3.2.10 Empstat

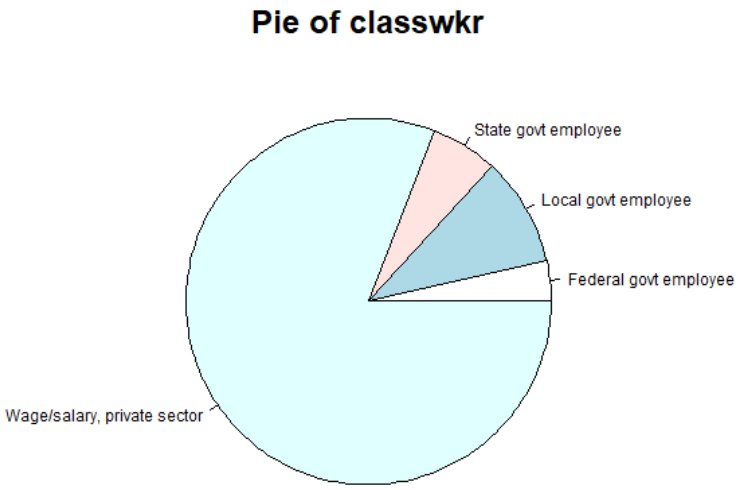


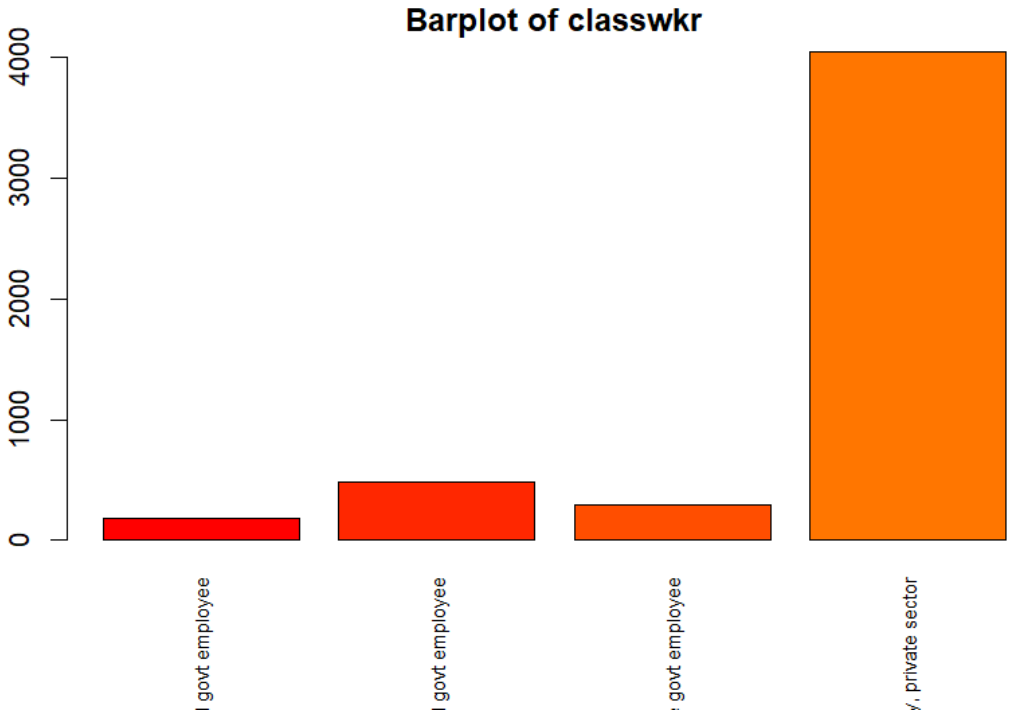


Description:

The graphs allow us to see that almost every individual in the dataset is currently working. Every individual has a job, but a small amount of them are not currently working.

3.2.11 Classwkr

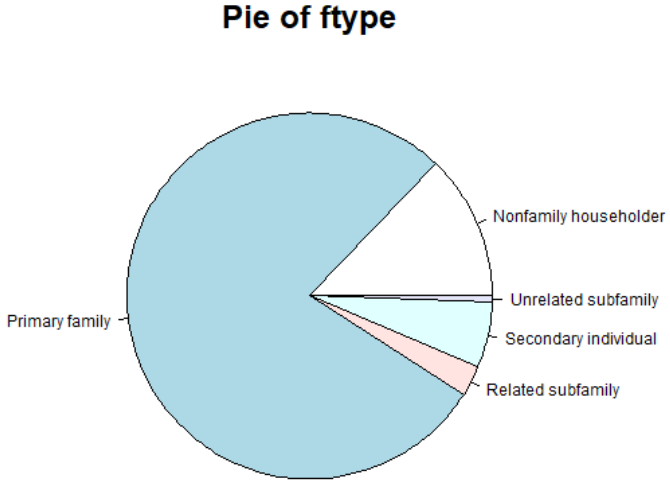


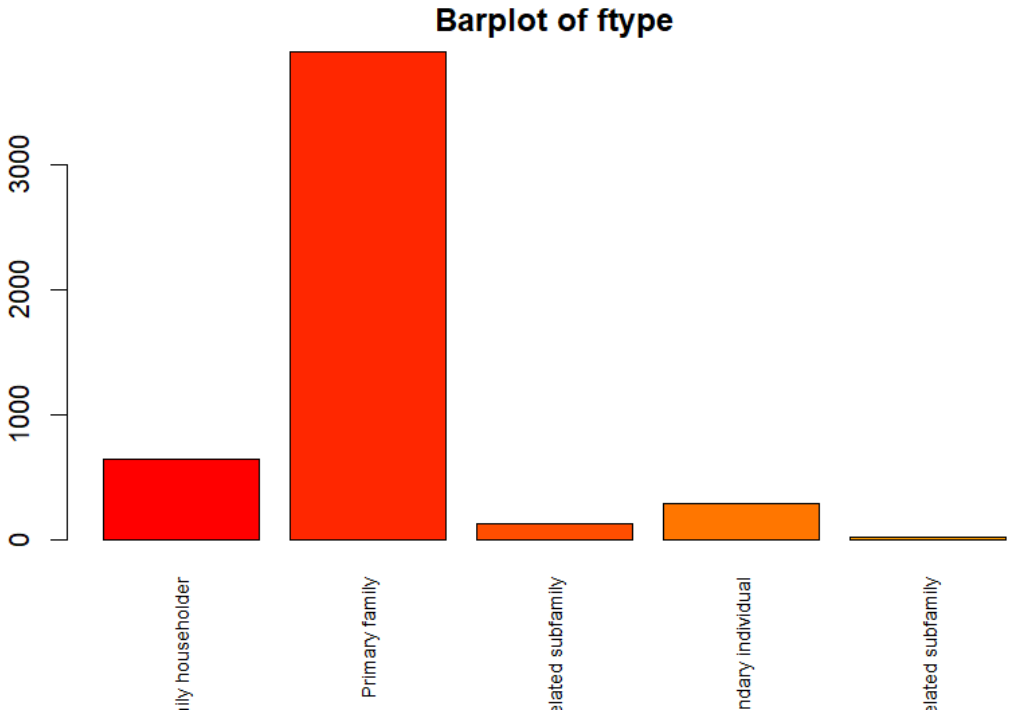


Description:

The graphs allow us to see that the most representative class of workers is from the “Private sector”, which represents more than two thirds of the whole dataset.

3.2.12 ftype

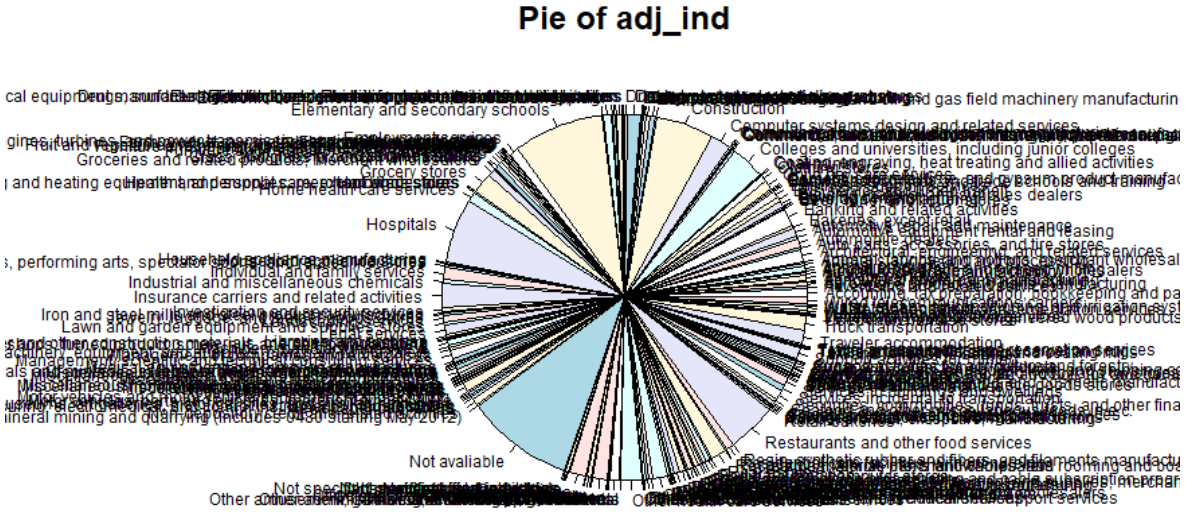


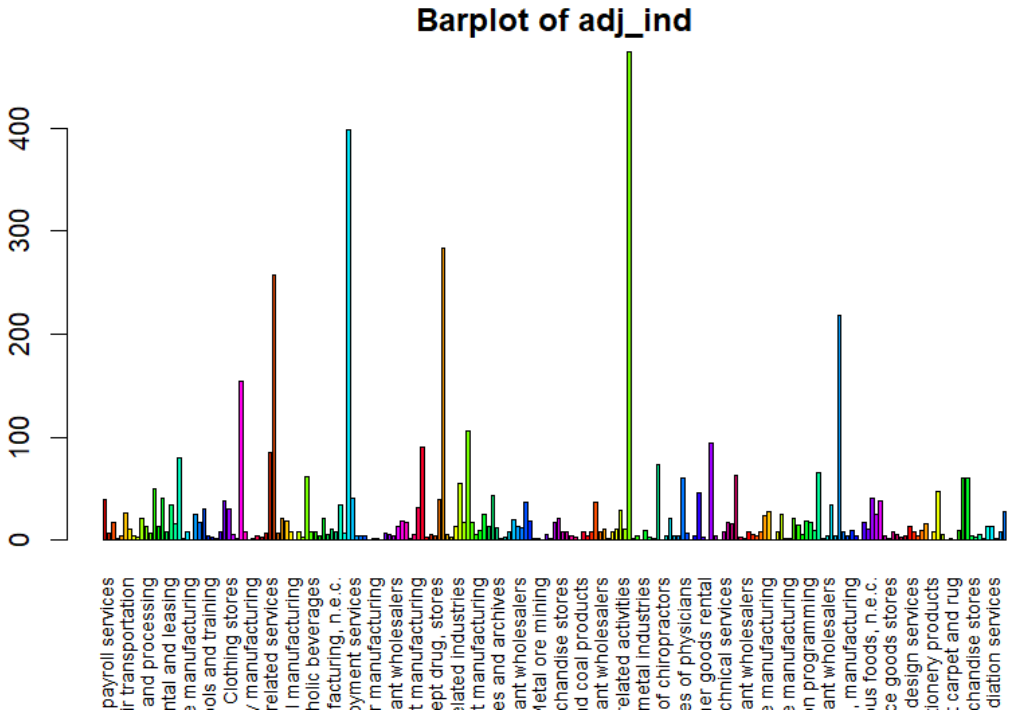


Description:

The graphs allow us to see that the most representative family type is “Primary family”, which represents two thirds of the whole dataset.

3.2.13 Adj_ind

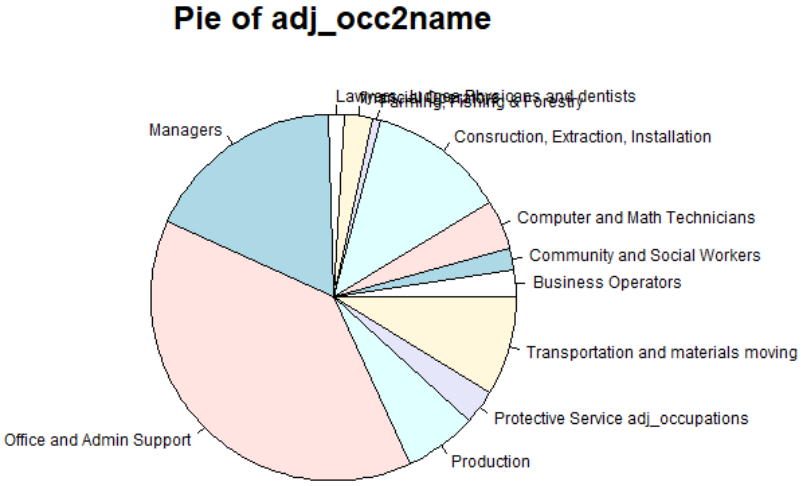


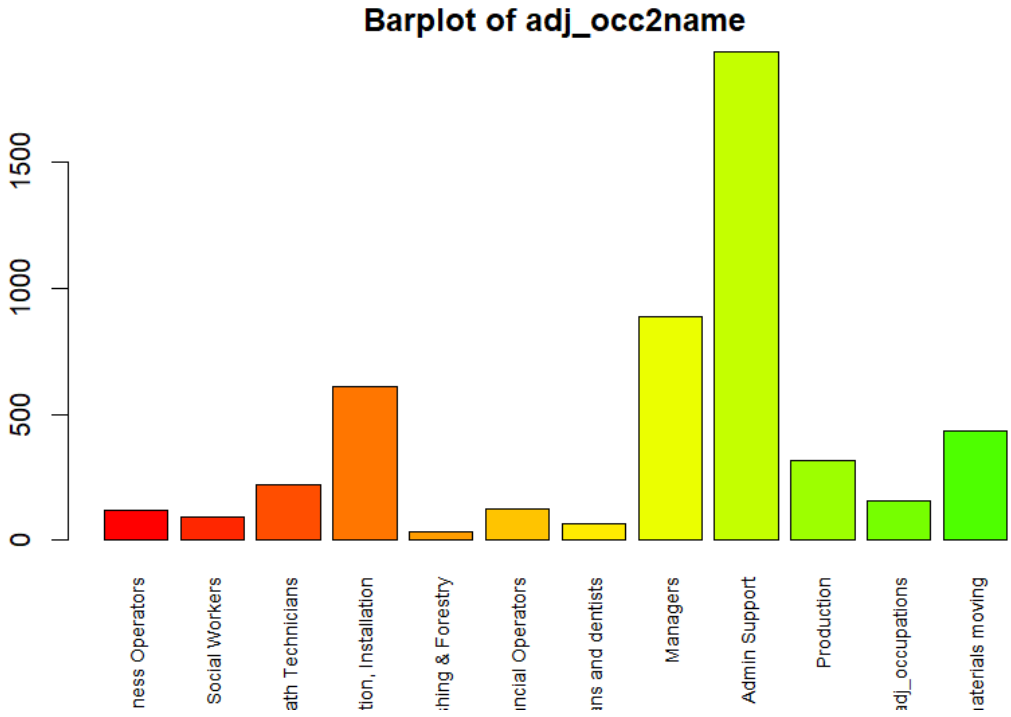


Description:

The graphs allow us to see that the most representative Industry fields are “Elementary and secondary schools”, “Hospital”, “Construction” and “Restaurant and other food services”. Nevertheless, the bigger sector is composed of people from which data are not available.

3.2.14 adj_occ2name





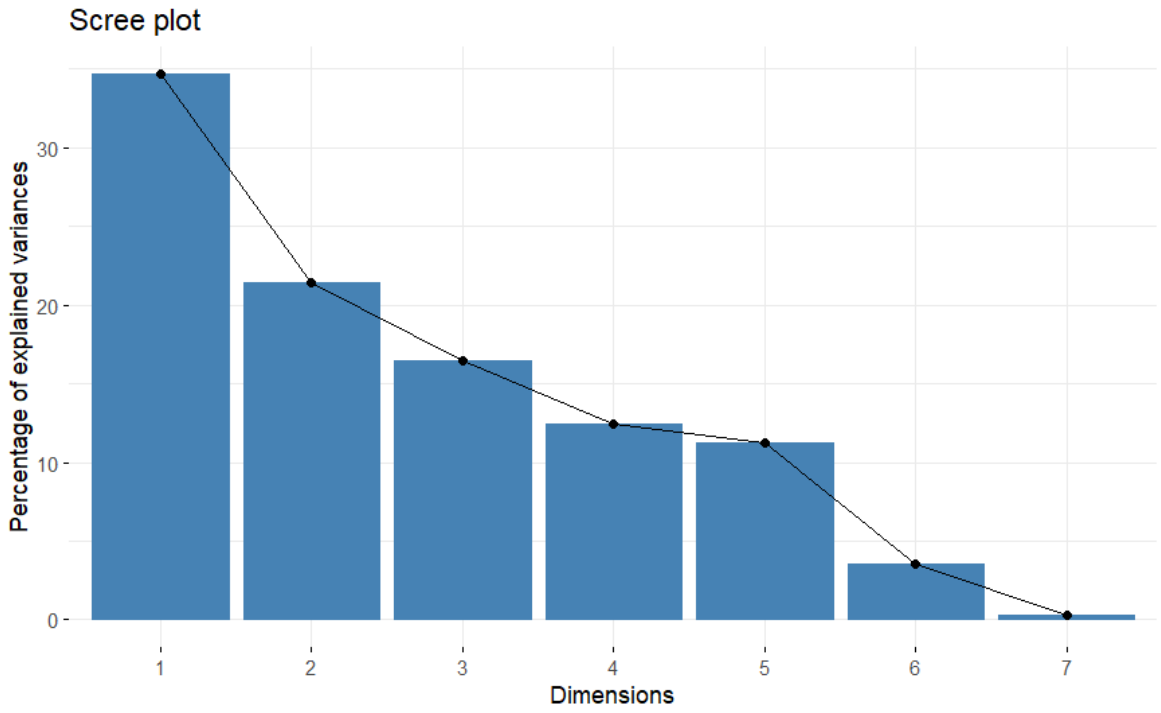
Description:

The graphs allow us to see that the most representative occupations are “Office and Admin support”, which represent the half of the dataset, “Managers” and “Construction, Extraction, Installation” are the other significant ones.

4. PCA

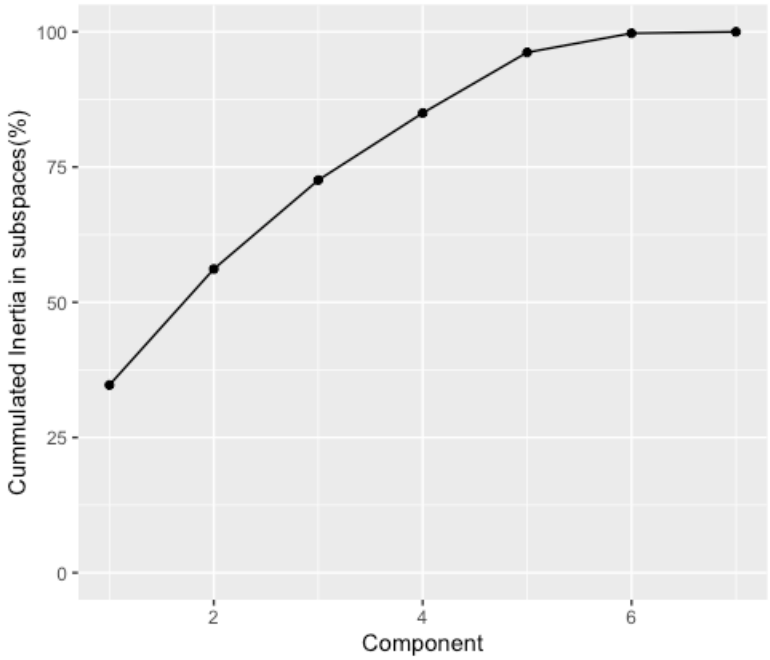
4.1. Scree plot and accumulated inertia

Firstly, we will see the principal component analysis of the continuous variables. To do that, we have created a new data frame only with the 7 numerical variables, with no missing values.



Description:

In this plot, we can see the total inertia represented in subspaces. The first component represents 34.70% of total inertia, the second 21.45%, the third 16.44% and the fourth 12.40%.



Description:

In this plot we can see the accumulated inertia represented in subspaces. With the third firsts components we obtain a 72.58% of the inertia, and adding the fourth component to them, we can obtain an 84.98% of the inertia.

So, we have selected 4 dimensions (84.98%) to keep the 80% of total inertia. The following data show the principal component values:

	PC1	PC2	PC3	PC4
numprec	0.03183951	0.04621265	-0.716909394	0.5411894845
age	-0.07952334	-0.19420880	0.663599292	0.4503005680
wkswork1	-0.35466931	0.29590017	0.112453621	0.5911003377
uhrswork	-0.51063621	0.25339074	-0.057464419	-0.3906883482
incwage	-0.43432937	-0.50718394	-0.125786057	0.0002949643
annhrs	-0.58007888	0.33095865	0.003516627	-0.0478728327
hrwage	-0.28458588	-0.66454073	-0.117874618	-0.0038013950

We have to analyze in detail the vectors that each component forms to interpret what type of information each of them contains. The values can range from -1 to 1. Values that are close to -1 or 1 indicate that the component is significantly affected by the variable, and values close to 0 indicate that the variable has little influence on the component.

For example, the third component is the result of the following linear combination of the numerical variables:

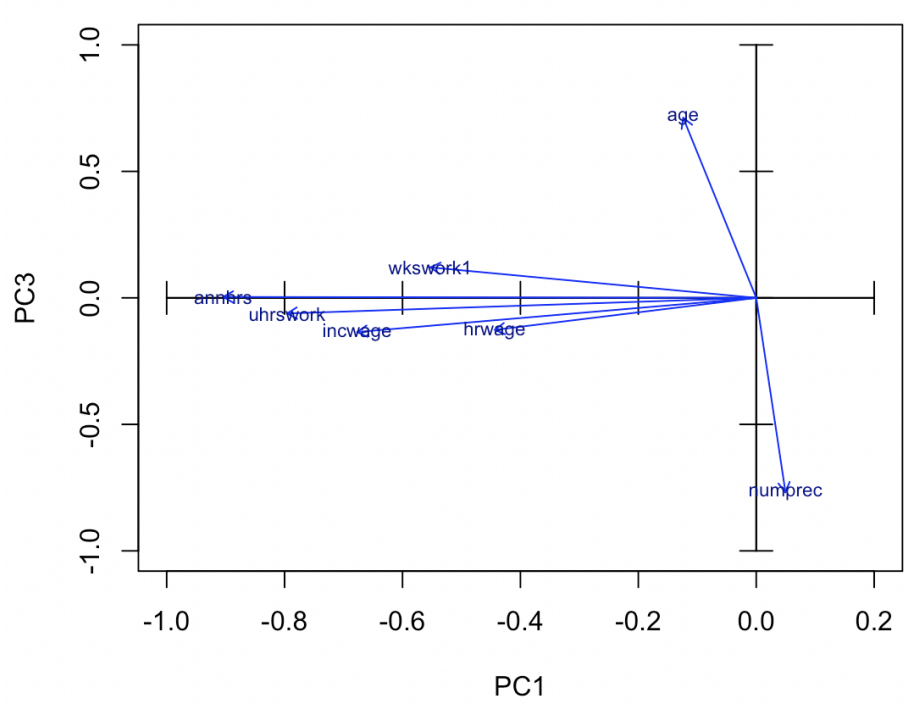
PC3 = - 0.717 numprec + 0.664 age + 0.112 wkswork1 - 0.057 uhrswork - 0.126 incwage + 0.004 annhrs - 0.118 hrwage

We can appreciate that the variables *numprec* and *age* have influence on the third principal component PC3.

Another example, in the first principal component PC1 the two variables mentioned above don't have influence on it, because the values are very close to 0.

4.2. Numeric variables

To do the analysis, we have selected the first and third principal components to represent all the following plots.



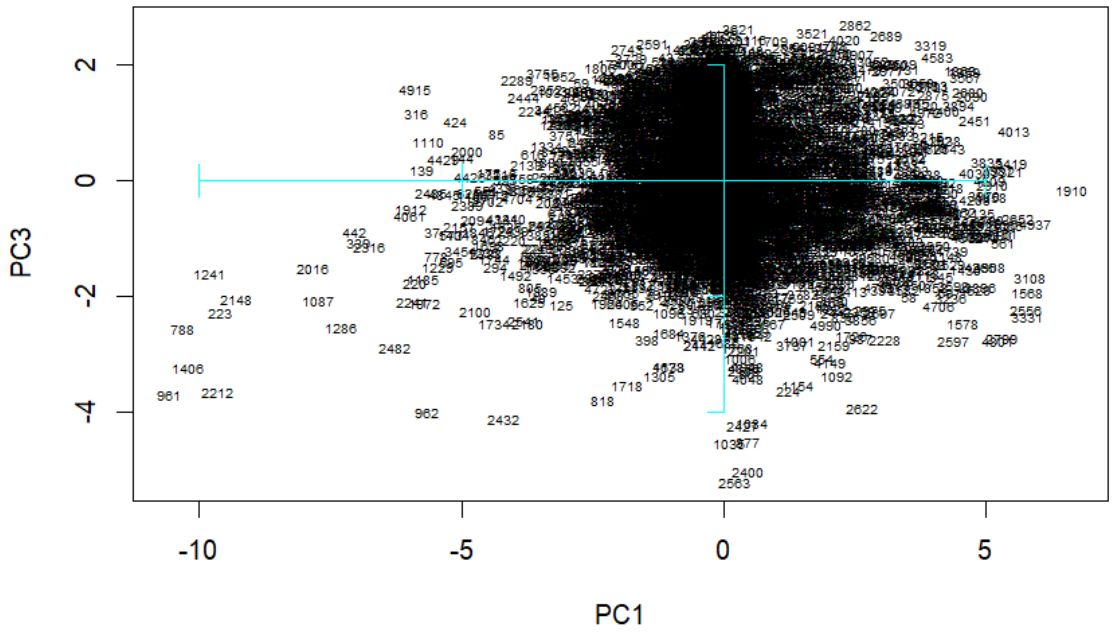
Description:

This plot represents the projection of numerical variables. As it has been mentioned previously, we can observe that the variables *age* and *numprec* have influence on the third principal component, because the angle between their projections and the axis that refers to the component is small. The other 5 variables have influence on the first component for the same reason, especially the variable *annhrs*, which makes an angle of almost 0 with the X axis (PC1).

Another thing to comment is the correlation between variables. The smaller the angle between the projection of two variables is, the more correlation they have. So, the variables *wkswork1, annhrs, uhrswork, incwage and hrwage* have relation between them, and no relation with variables *age* and *numprec*. These last two have a negative correlation between them.

The PC1 seems to describe the relation among time and money, related to the time spent and the amount of money generated. While in PC3 seems to take condition about the lifetime of people and the quantity that stand in a same place living together.

4.3 Individual's plot

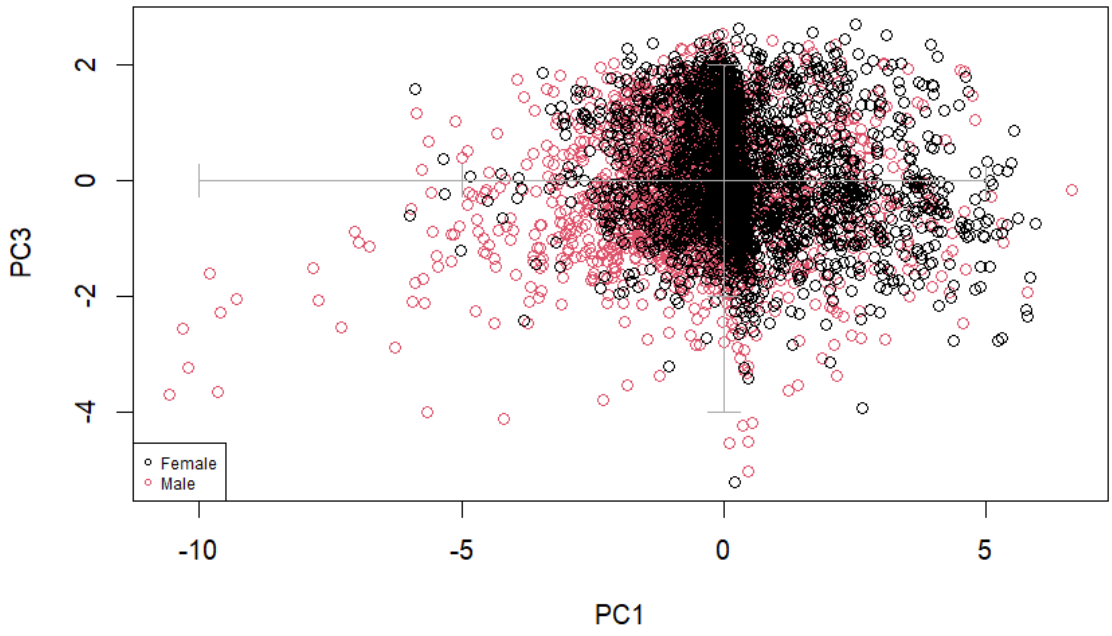


Description:

This is the plot of individuals. That means that we can see, for each person of the dataset, the relation between PC1 and PC3 that they have. Taking a global view of the plot, we can appreciate that there is a big spot in the center, so we can relate it to having a zero variance in general terms. We could draw a circle on the spot having 2 of radius, so the majority should be included there. So we can conclude that there is a meaning less between the relation of each individual and both PCs.

4.4. Qualitative variables

Secondly, we will see plots that correspond to the modalities of qualitative variables.

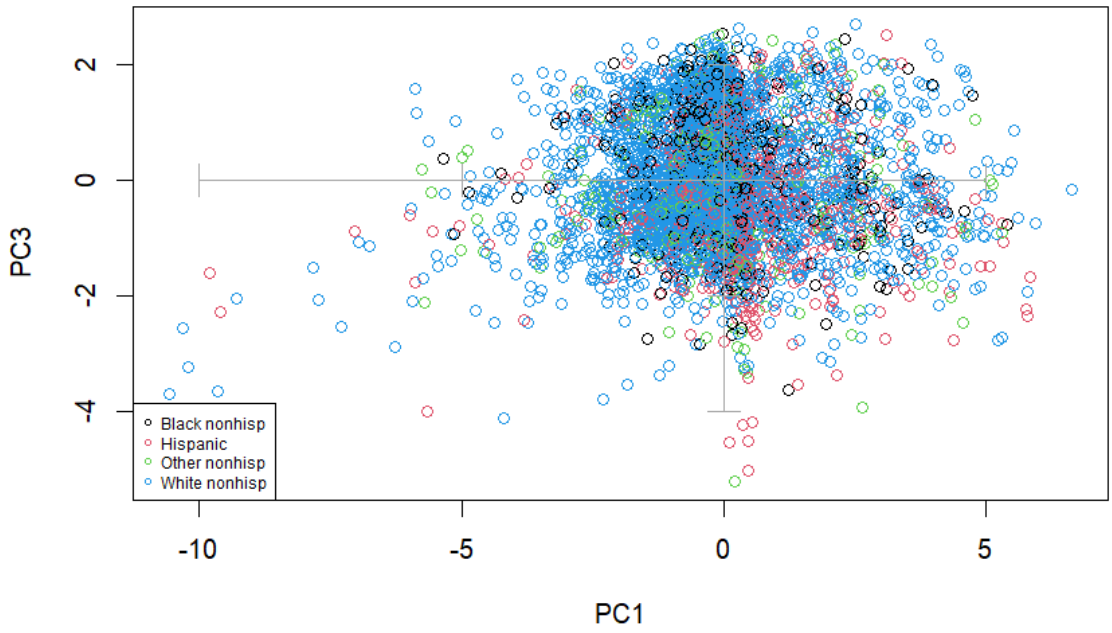


Description:

This is the plot of individuals differentiating the qualitative variable *sex*. It has 2 possible modalities: female and male. We can observe that neither of them predominate, there is 50% of women and 50% of men in the data.

We can define, taking a wide perspective, that the covariance of the plot has tendency to zero, which means that there is just a little relation.

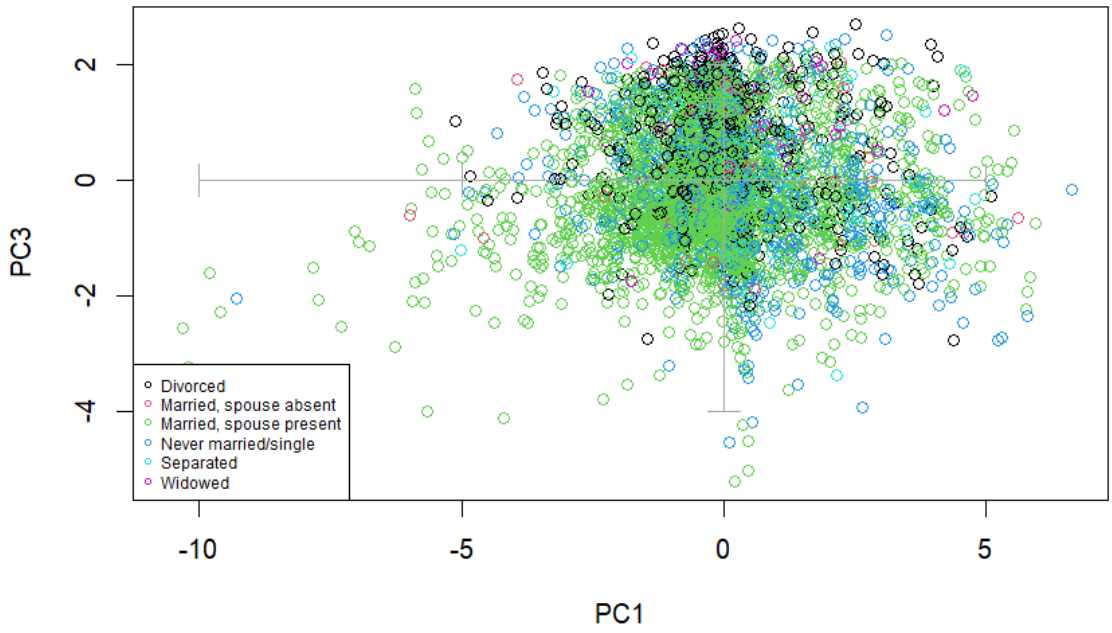
It's difficult to obtain a conclusion, in terms of accuracy, but in general seems to be that, as covariance shows, both sex haven't a high relation with PC1 and PC3. The first one referred to the relation between money and time, and the second one to the lifetime and amount of people living daily around in a particular space. But if we take a deep look, we can appreciate that there are more pink points in first and third quadrant and more black points in second and fourth ones. In terms of the PC1 we can consider that there is a higher benefit to male respect female. In terms o PC3 there are fewer differences between them.



Description:

This is the plot of individuals differentiating the qualitative variable *race*. It has 4 modalities: black nonhisp, hispanic, other nonhisp and white. We can see that the data has more white people than people from other races.

Considering the white race as the predominant in the plot because of the difference of quantity between other races, we can consider, by having a wide perspective, that the covariance is zero, so seems that the relation is insignificant on both components. We can apply the same reasoning to hispanic and black nonhisp. Moreover, focusing on hispanic, there is a huge number of them, closer to PC3 on the fourth quadrant, that seems to considerate a younger/medium age.



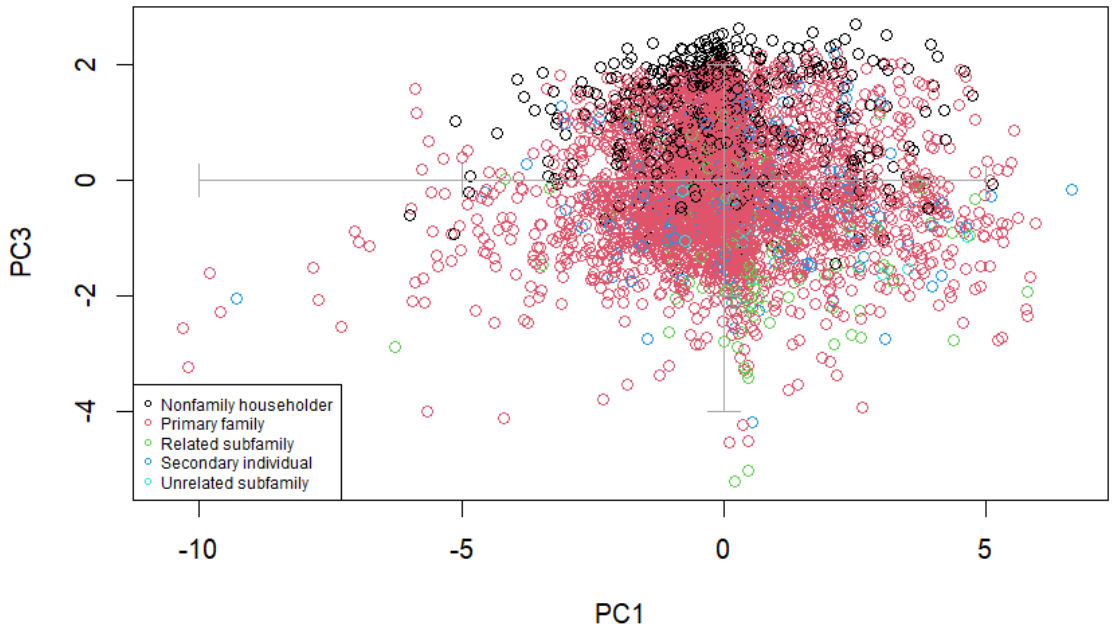
Description:

This is the plot of individuals differentiating the qualitative variable *marst*. It has 6 modalities: divorced, married spouse absent, married spouse present, never married/single, separated and widowed. We can see that the modality with more individuals is “married, spouse present”, and the second with more individuals is “never married/single”. The other modalities have less representation in the data.

In general terms, we can see that there are three big groups: the married (spouse present), the separated and the divorced. For the first one, that can be considered the biggest, in terms of quantity, in respect to other categories, we can consider having a zero covariance because they are focused near the zero and the two range. That means that there isn’t a significant relation in respect to both dimensions.

In the second case, we can appreciate a similar zero covariance, but the focus could be related from zero to three. Again it has a meaningless in relation of both dimensions.

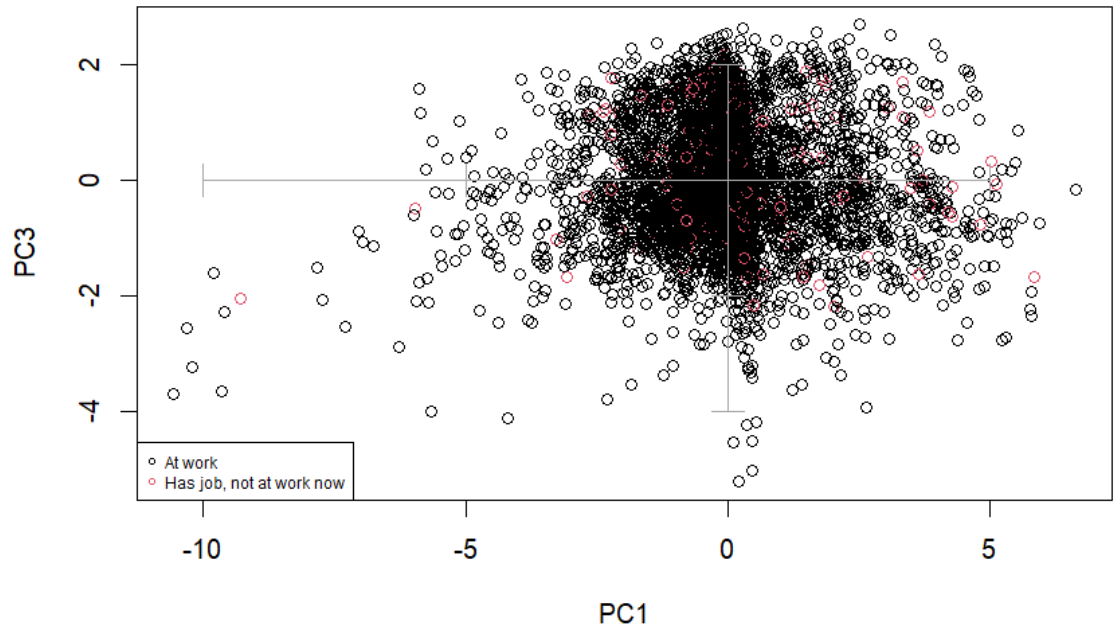
Finally, in the third case, we can look that there is a huge amount of points concentrates in the first and second quadrant and near to the PC3. So we can conclude that this group take part from the medium/high age of people.



Description:

This is the plot of individuals differentiating the qualitative variable relate. It has 5 modalities: Nonfamily householder, primary family, related subfamily, secondary individual and unrelated subfamily. We can see two main groups, primary family and nonfamily householder.

On one hand, for the first group, we can see that covariance is zero, or are related to it because painting a circle of radius 2, the majority of points are included. So we can say that the relation isn't significant between this category and both dimensions. In the other hand, for the second group, we can appreciate a majority distribution of the points in the first and second quadrant. So we can deduce that this category is mostly composed by medium/high age.



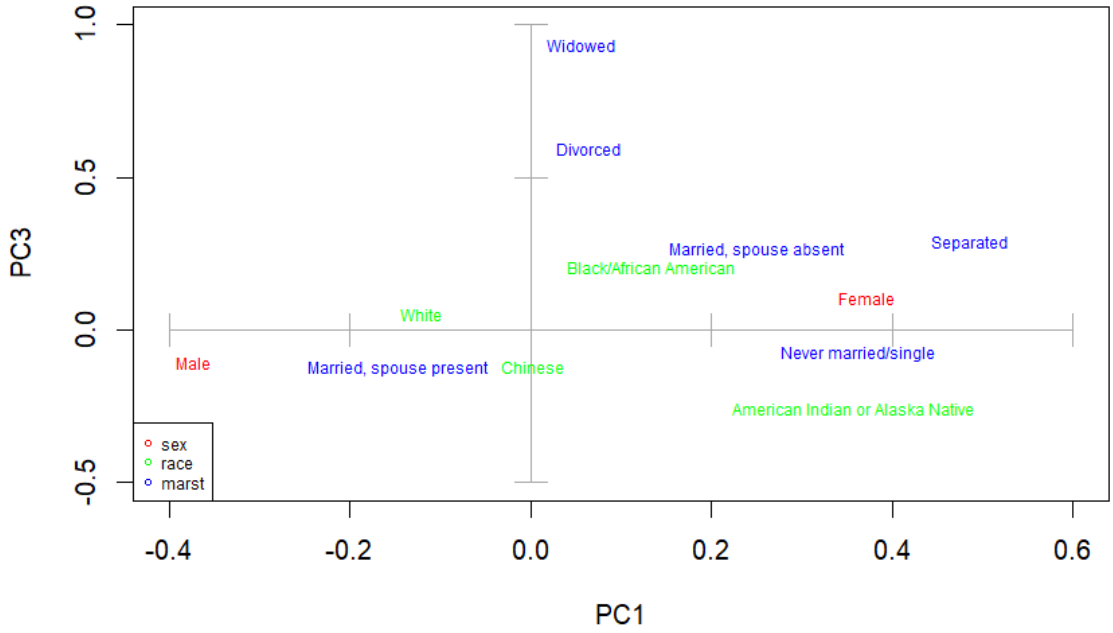
Description:

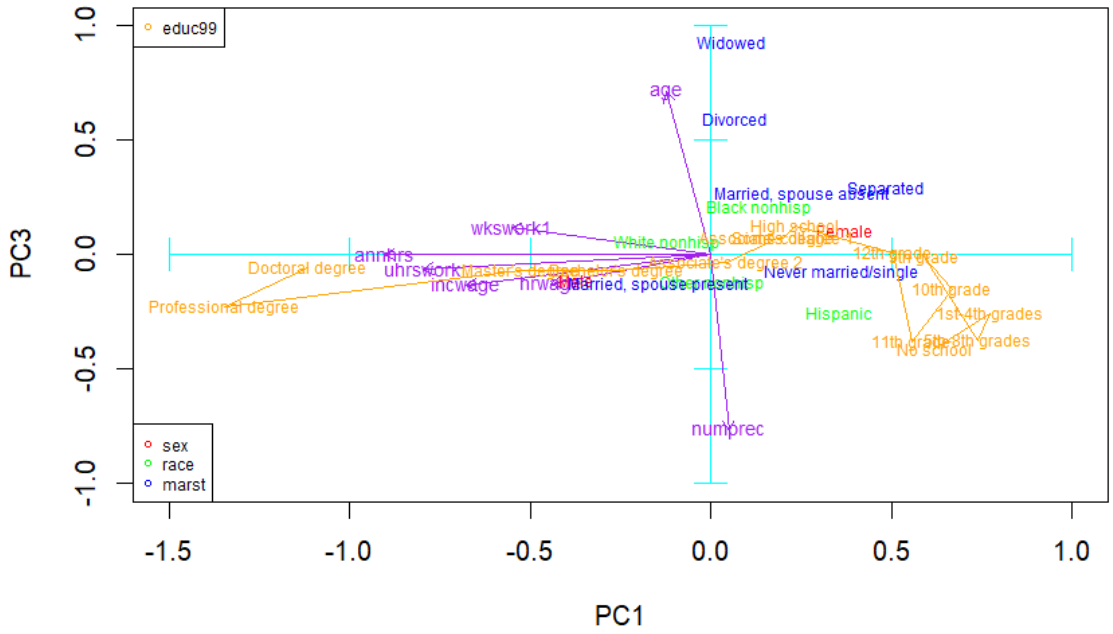
This is the plot of individuals differentiating the qualitative variable *empstat*. It has 2 possible modalities: at work and has job, not at work now. We can observe that there is a big difference between them in the number of individuals. People who have work represent more than the 80% in the data.

Finally, to represent all qualitative variables together, we have had to divide them in several groups because joint representation saturates.

In the following graphs, we can see the representation of the categorical variables *sex*, *race* and *marst*, as we can see in the legend. In the second graph we have added the projections of numerical variables in background, and an ordinal qualitative variable called *educ99*. This variable has 15 levels, which indicate the educational attainment grade.

4.5. Numeric and qualitative variables


PC3
PC1
1.0
0.5
0.0
-0.5
-0.4
-0.2
0.0
0.2
0.4
0.6
Widowed
Divorced
Married, spouse absent
Separated
Black/African American
Female
White
Never married/single
Male
Married, spouse present
Chinese
American Indian or Alaska Native
sex
race
marst


PC3
PC1
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
educ99
Widowed
age
Divorced
Married, spouse absent
Separated
Black nonhisp
wkswork1
Female
White nonhisp
annhrs
Doctoral degree
uhrswork
Never married/single
incwage
Married, spouse present
10th grade
Professional degree
Hispanic
1st-4th grades
numprec
sex
race
marst

Description:

From both plots, we can conclude the following:

- People who own a higher degree will earn more money compared to those who are only high-school or lower educated.

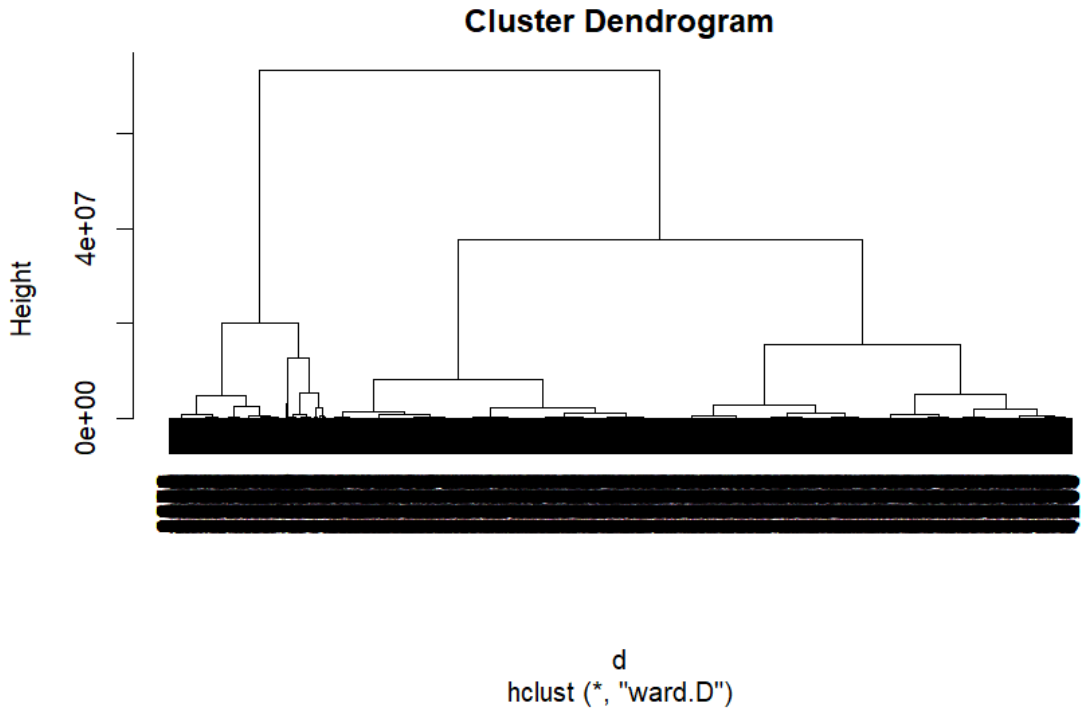
- If you gather these three characteristics: male, white and married, you are more likely to earn money than other individuals who don't have these characteristics.

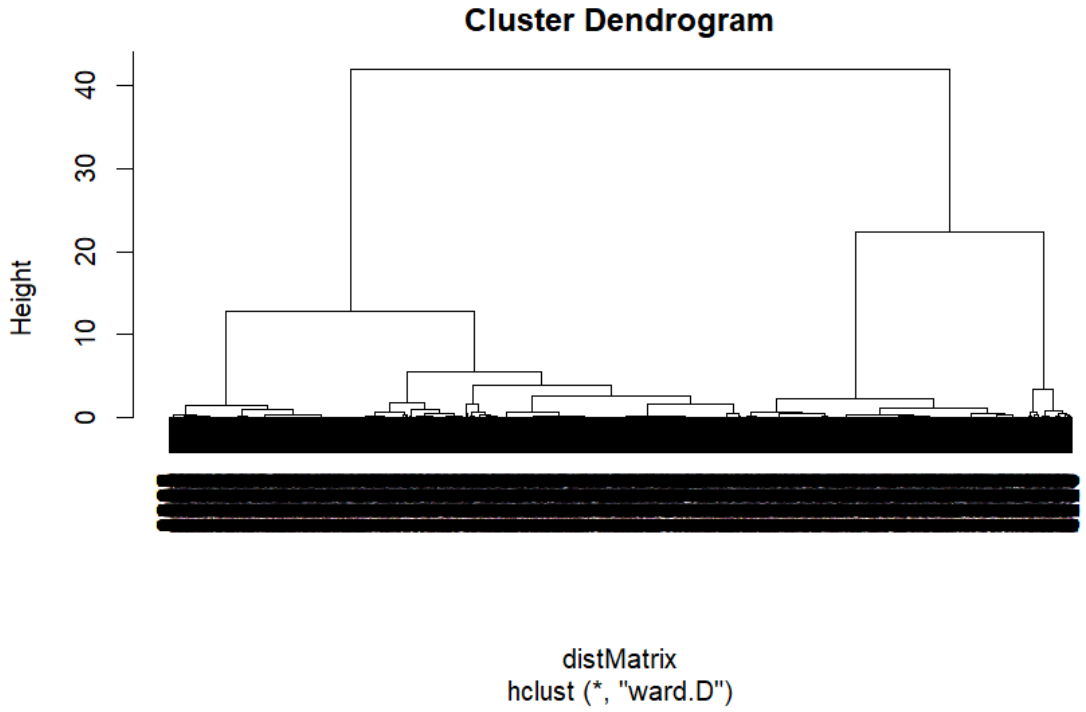
- The widowed and divorced are more likely to be older.

5. Clustering

In this part, we will have a look at how we can group our data, so we can reach different clusters and see from them the classes that can be produced. Firstly, we have used hierarchical clustering to decide how many clusters we should consider organizing the dataset the best way possible.

5.1. Dendrogram consider organizing



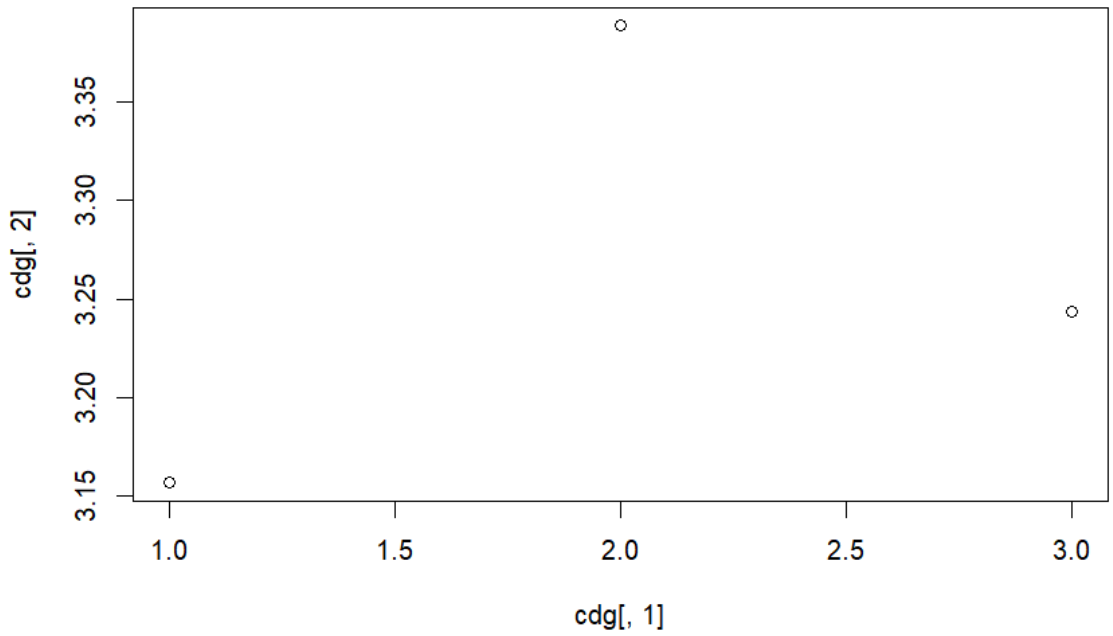


Description:

This is our dendrogram, obtained from our dataset. We have decided to cut it at the first height, so we have 3 different clusters. As we can see in the picture, we have considered cutting in the first height because we can appreciate three main clusters.

The differences between these two plots is that the second is focusing in a deeper part of the general dendrogram seen above. Again, we can appreciate cutting the second tree in the second height and obtaining three main clusters.

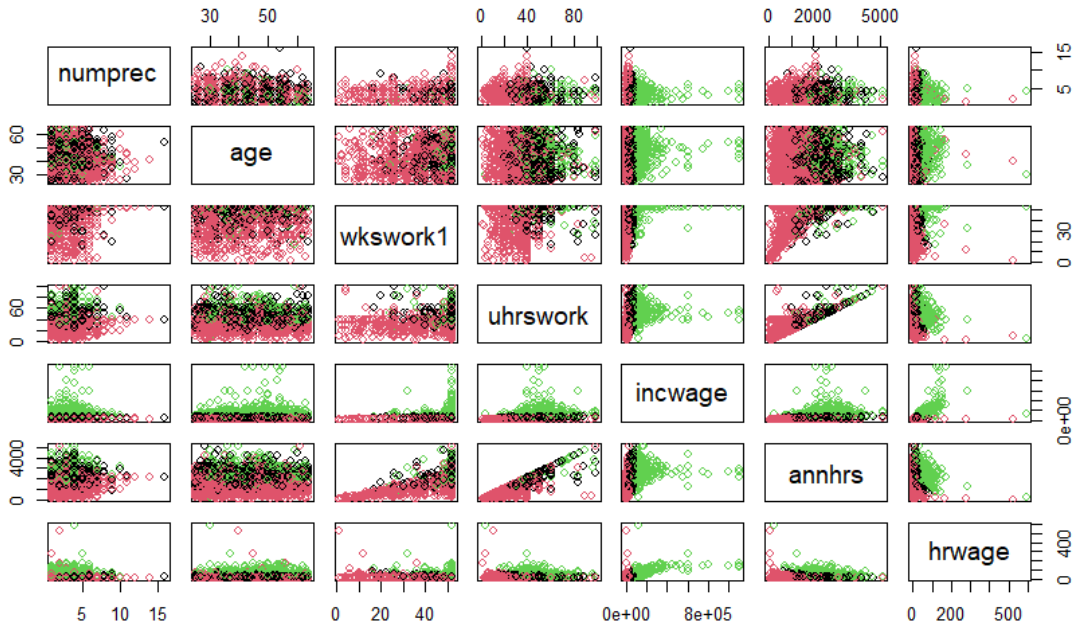
5.2. Kmeans



Description:

Here we can see, by using kmeans, the three centroids of the three clusters explained above. Clearly, we can observe the distribution of the different groups. Moreover, the number of individuals on each cluster are: first cluster (2236), second cluster (1889), third cluster (875).

5.3. Partition visualizer



Description:

In the matrix of above, we can appreciate the plots generated between the different numerical values when the clusters are introduced. So we can observe, in the majority of them, with an easy view, three groups differed by colors (pink, black and green) corresponding to the main groups.

In this section, we will focus on the main important plots from the picture showed before.

Description:

In these plots, we can see the relation between the wage and salary income value for a determinate age. Looking at them, we can observe the three groups mentioned before. In general terms, we can appreciate that all the classes are under the 100k incwage, except a few points from the green one.

Moreover, while the pink and black ones seem to maintain a similarity in their values depending on the age, on the green ones we can see some heights for the middle age users, around (35-55).

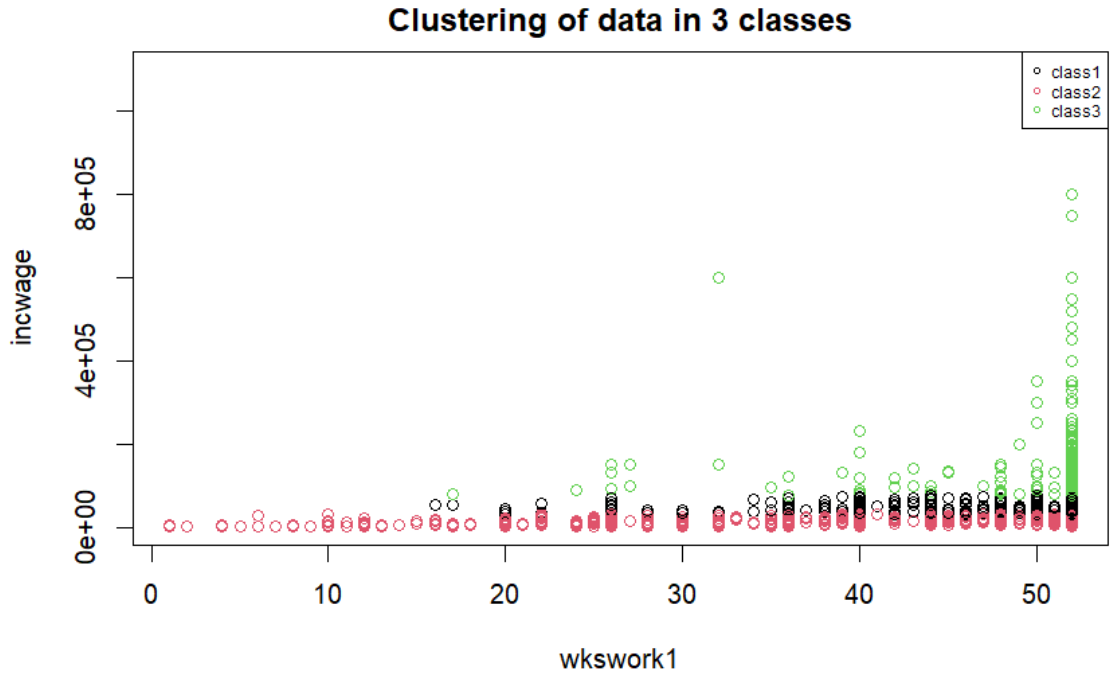
As a conclusion, the classes show the acquisition status related to the monetary possession and how they are related among the age.

Description:

In these plots, we can see the relation of the hourly wage for a specific age in a user. Looking at them, we can see again the main three groups. In general terms, we can appreciate that all of them are under the 100 hourly wage, there are few points corresponding to the green class and some of the pink ones that get out of the previous range.

Moreover, we can observe similar values between pink points while in the black group there are some differences related to the hourly wage. Also with the green class, where the difference is even bigger. Therefore, there is a height tendency in the green class, in the middle age (35-55), which increases the hourly wage.

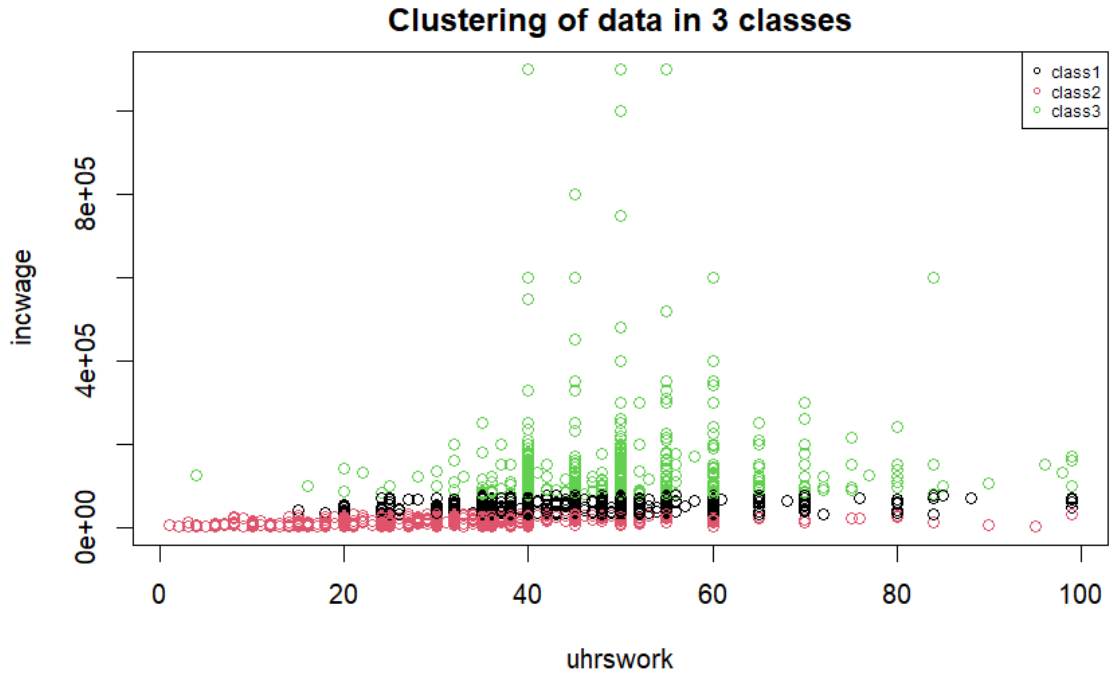
.



Description:

In these pictures, we can see the relation between the wage and salary income and the weeks worked in the last year. We can see that the majority of the users that had worked for 50 weeks during the last year, have the highest wage and salary income. Also, we can appreciate that the second class, which is formed by individuals with a smaller salary, has individuals working only more than 15 weeks a year, that is normal, although it has an important number of individuals working a significant number of weeks too. There are individuals from the third class, with the highest salary, that have received more money working less weeks than people from the first or second class working more weeks a year.

As a conclusion, we can summarize that individuals with the higher salary have worked a lot of weeks a year, but working the most weeks doesn't relate to having the highest salary.

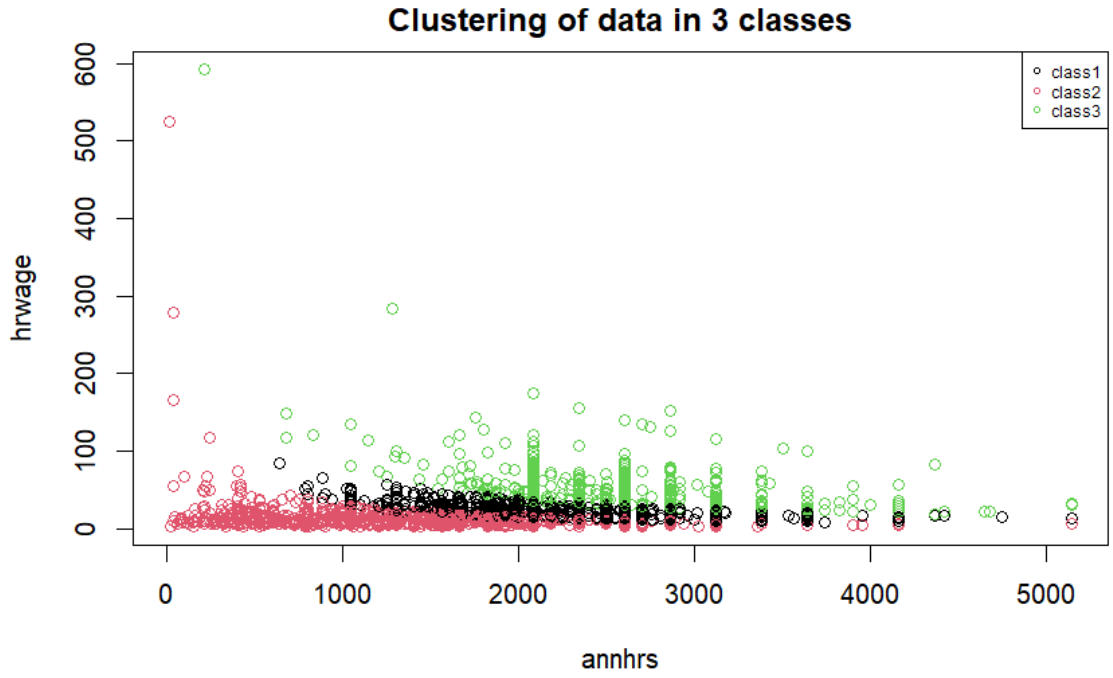


Description:

In these plots, we can see the relation between salary and the usual hours of work per week. The three main clusters represented show us different classes on generating the total amount of money and the time in realizing it.

For the pink points, we appreciate a focus on working between 20 and 40 weekly hours, and they have the lowest salary in relation with classes. In the black point case, we can observe a considerable spot between 36 and 56 weekly hours, obtaining much more wage in respect to the previous one. Finally, for green points, we can see that people working between 40 and 60 hours in a week have the highest wage and salary income.

As a conclusion, we can see that the relation between the weekly hours worked and the respect to salary is based on the amount of hours worked, which means that working more hours a person can generate more quantity, and also the profession valorization, because depending on it, the salary will be higher or lower.



Description:

These plots represent the relation between the annual hours worked and the hourly wage. We can see that the groups in these plots follow a log tendency, which means less annual hours worked, more hourly wage, and the opposite.

Individuals with the highest hourly wage are from the green class and they have a medium-high number of annual hours worked, expecting a few pink points that also have a high hourly wage but a low number of annual hours worked.

For the pink points, we appreciate a focus on working between 0 and 2000 annual hours, and the majority of them (excepting the few points commented before) have the lowest salary in relation with classes. In the black point case, we can observe a considerable spot between 1000 and 3000 annual working hours, obtaining much more wage in respect to the previous one. Finally, for green points, we can see that people working between 2000 and 2500 annual hours have the highest hourly wage.

5.4 PCA and Clustering conclusions

Based on the analysis generated in PCA, we have taken some conclusions about the information generated from the data. Focusing on the gender salary gap, we can appreciate that gathering these three characteristics (male, white and married), a person is more likely to earn more money beside people that don't have these. These three characteristics seem to be more considered by society policies. Moreover, we have observed that owning a high degree means to acquire a better position in society, so they can obtain job offers with a better wage. Furthermore, we have seen that there is a higher benefit in terms of money wage for males beside females.

In the PCA projection of numerical variables, we have seen that there is a correlation between the salary and the working hours, and there is no relation for example between age and salary. In the clustering part we have proved that in the clustering plot that represents salary and hours worked we can see a clear tendency, while in the clustering plot that represents salary and age, we can not see any, because the variables don't have any relation. So, we can conclude that variables with correlation produce tendencies in the clustering plots.

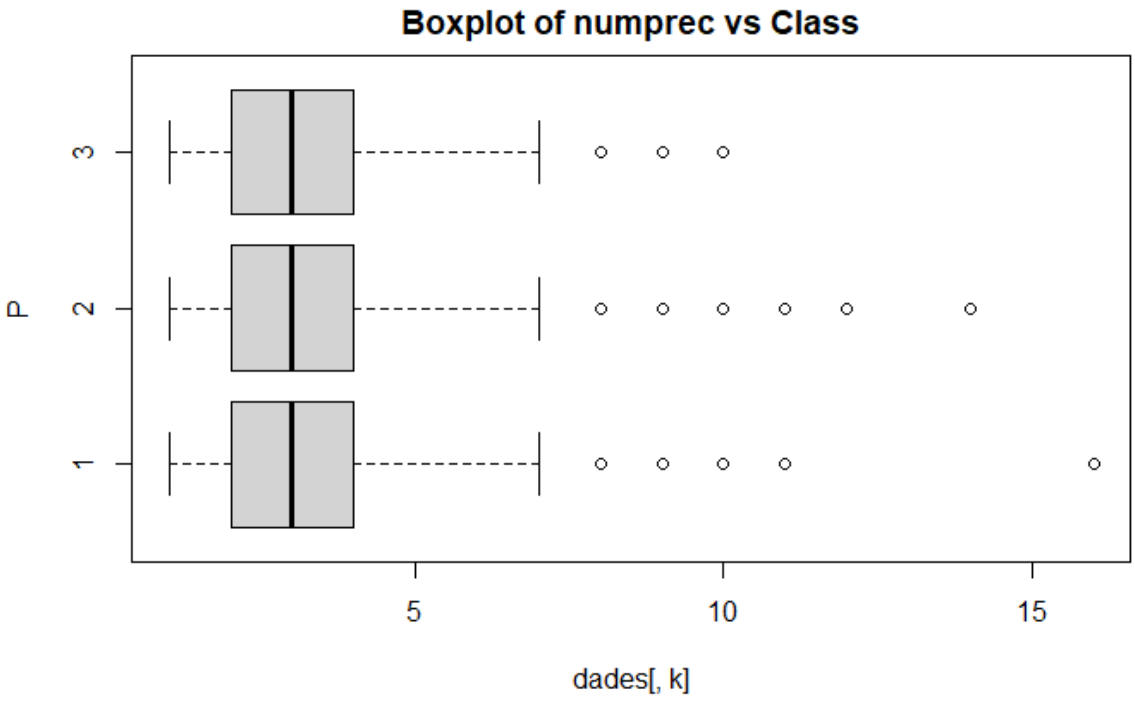
6. Profiling

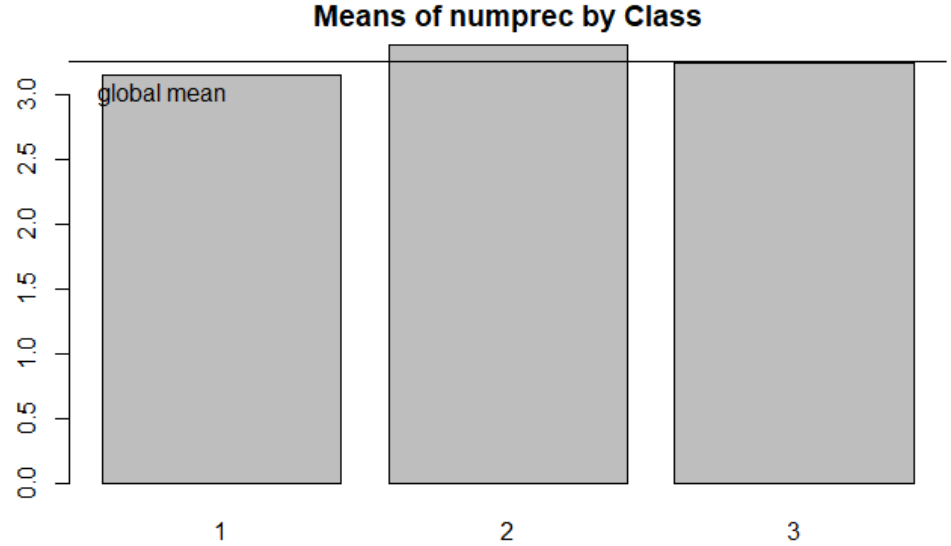
In this section we will present the outcomes of the profile based on hierarchical clustering. We will first present the plots for each variable, and then we will write down a cluster characterization in which we will highlight our results for each cluster based on the plots.

During this initial analysis we have talked about each cluster mentioning them by their numbers, but at the end, when making the conclusions about the results we obtained, we have given each of them a name.

6.1 Numerical variables

6.1.1 Numprec



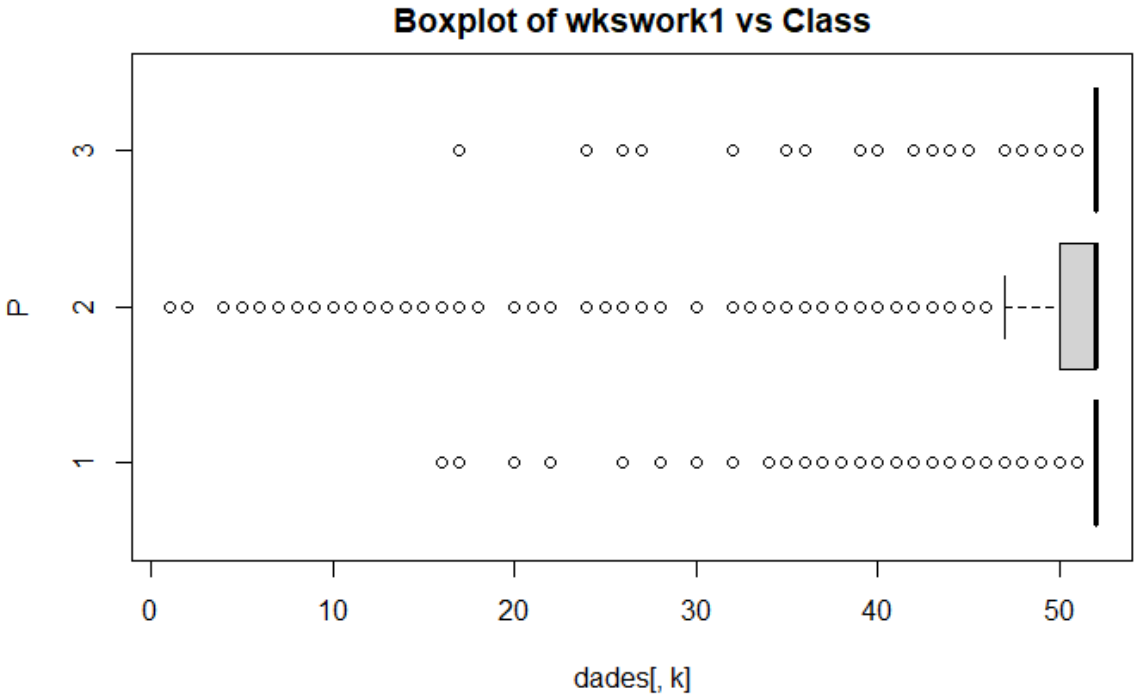


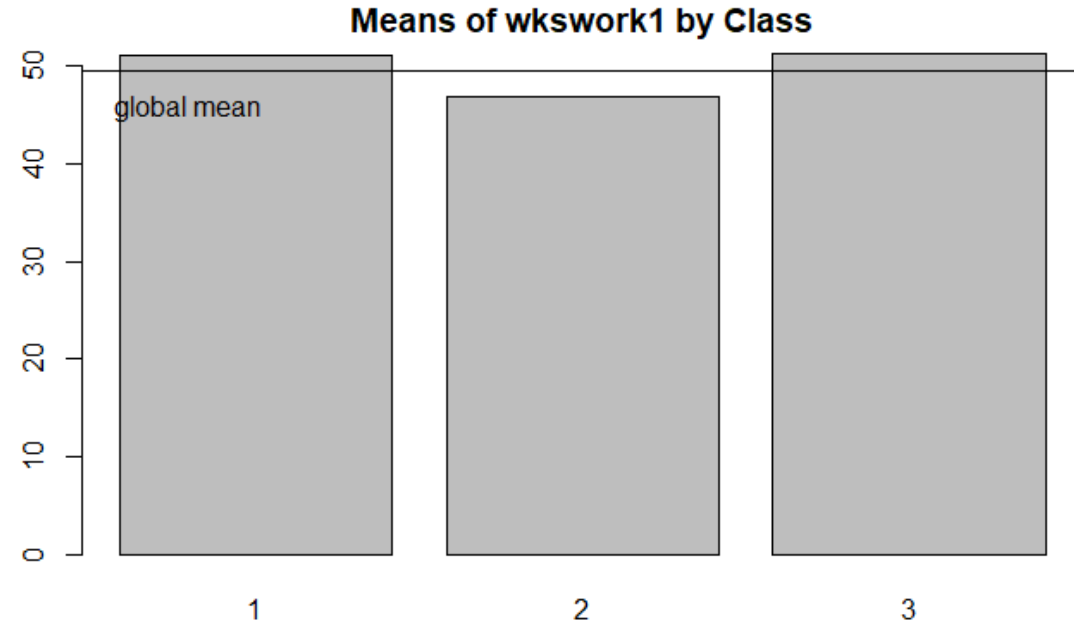
Based on these plots, we see that the variable numprec doesn't give us information about the clusters, since the three of them have almost the same means. The first cluster has the lowest mean by a small difference, but it's also the one with the smallest number of outliers, which is probably the cause of that.

6.1.2 Age

For cluster one and two, the means of the ages are almost the same. For cluster three as we can see it is slightly higher. At first glance it does not seem like a big difference maker.

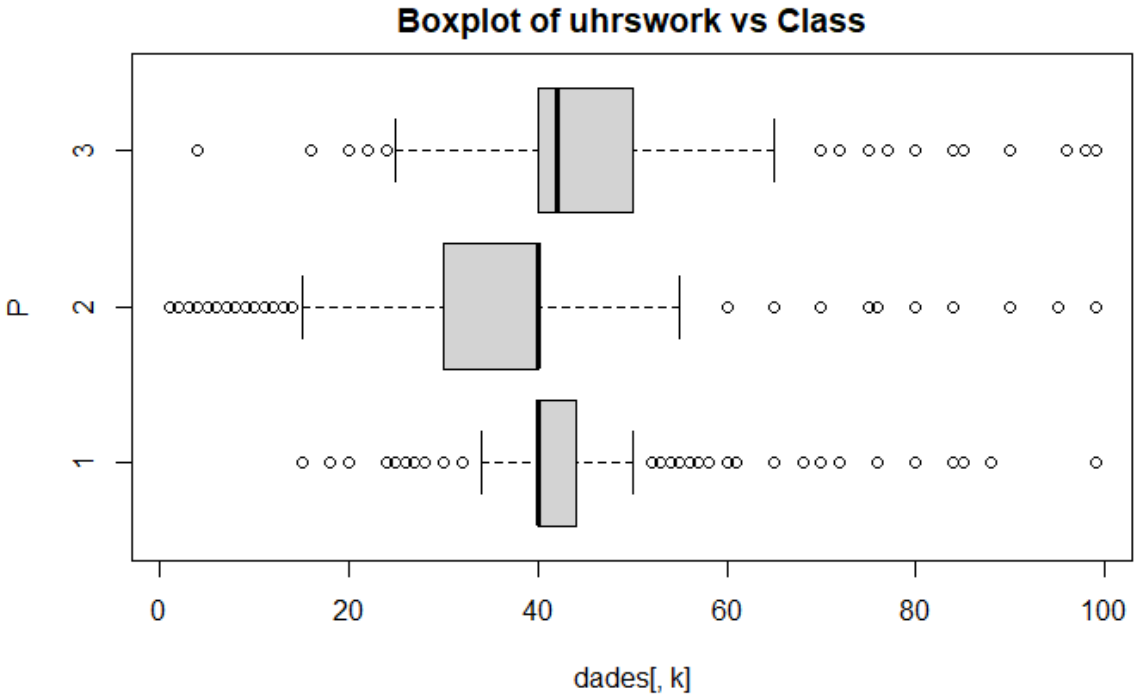
6.1.3 Wkswork1

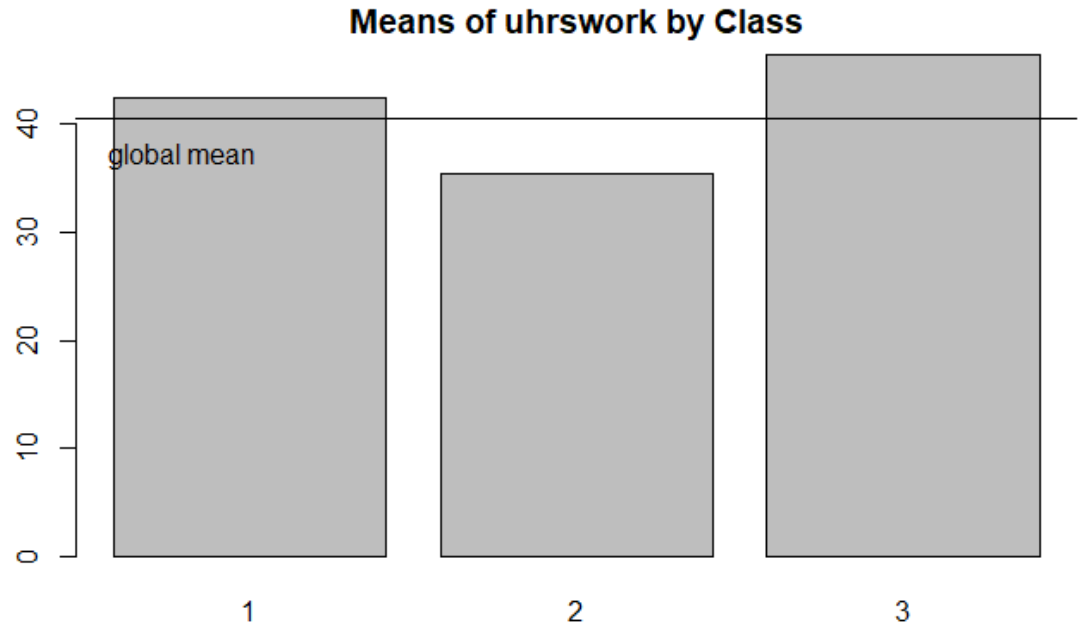




The three clusters have very similar means, the one for cluster two is lower and brings the global mean down. That is most likely due to the fact that they have the biggest amount of outliers. Although the weeks worked from cluster two are slightly lower, it doesn't seem like a huge difference.

6.1.4 Uhrswork

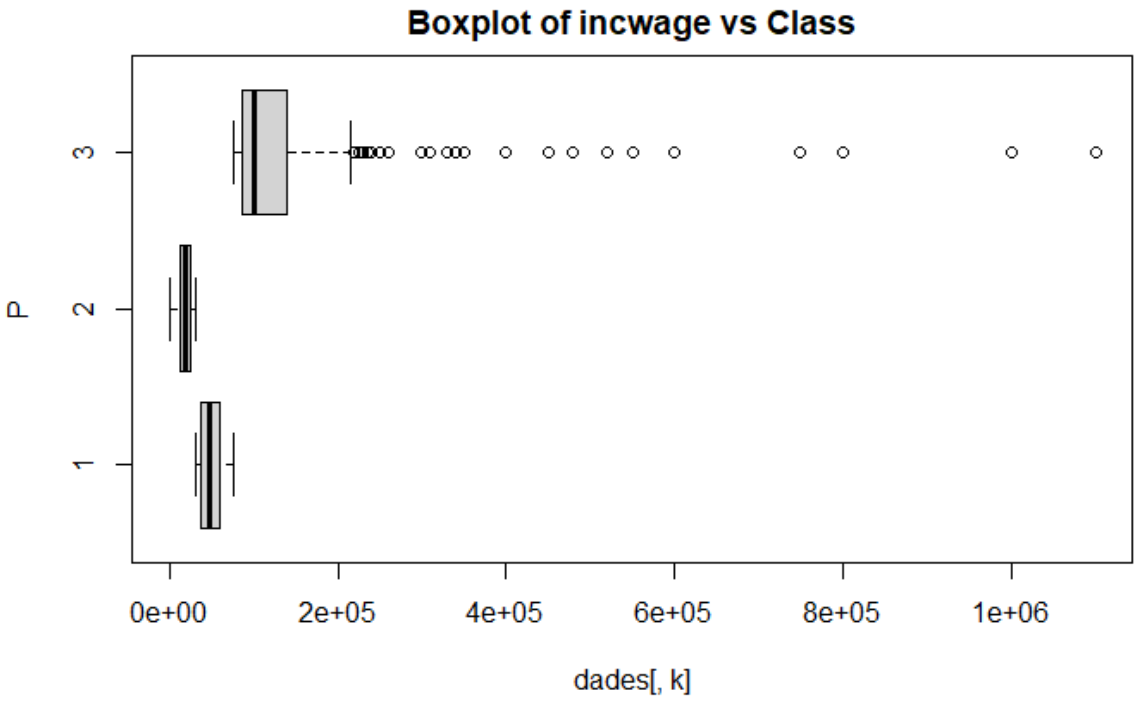


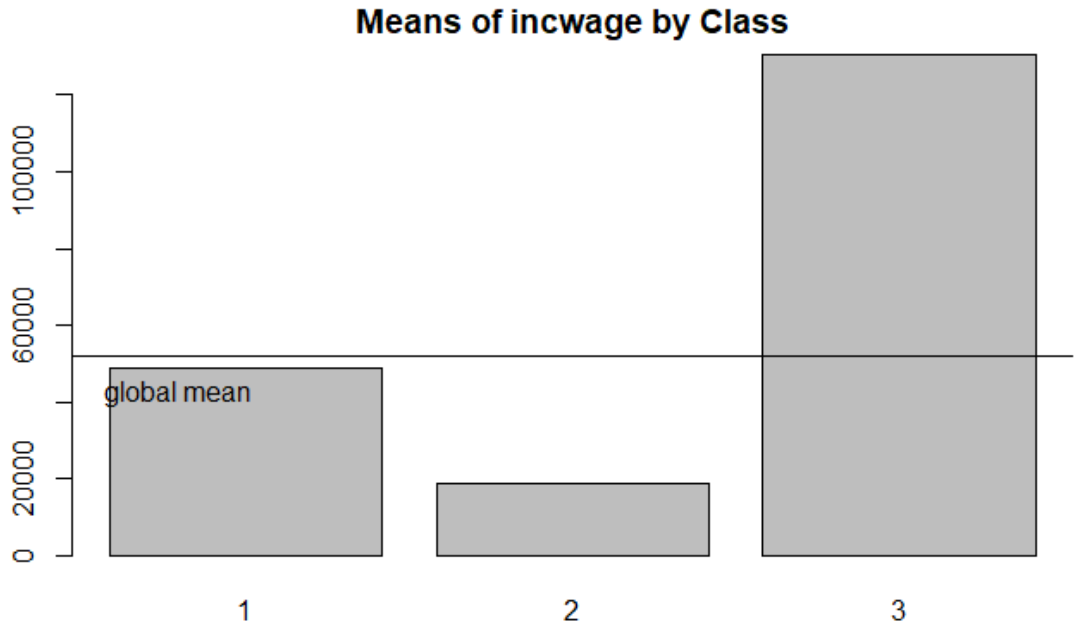


In this case again cluster two is the one with the lowest mean compared to the rest of clusters, and it's bringing the global mean down. In this case the difference feels slightly bigger proportionally than the ones we saw before.

The three clusters have a considerable number of outliers, the ones on cluster two have the lowest values, so they could be a reason why cluster two's mean is lower.

6.1.5 Incwage

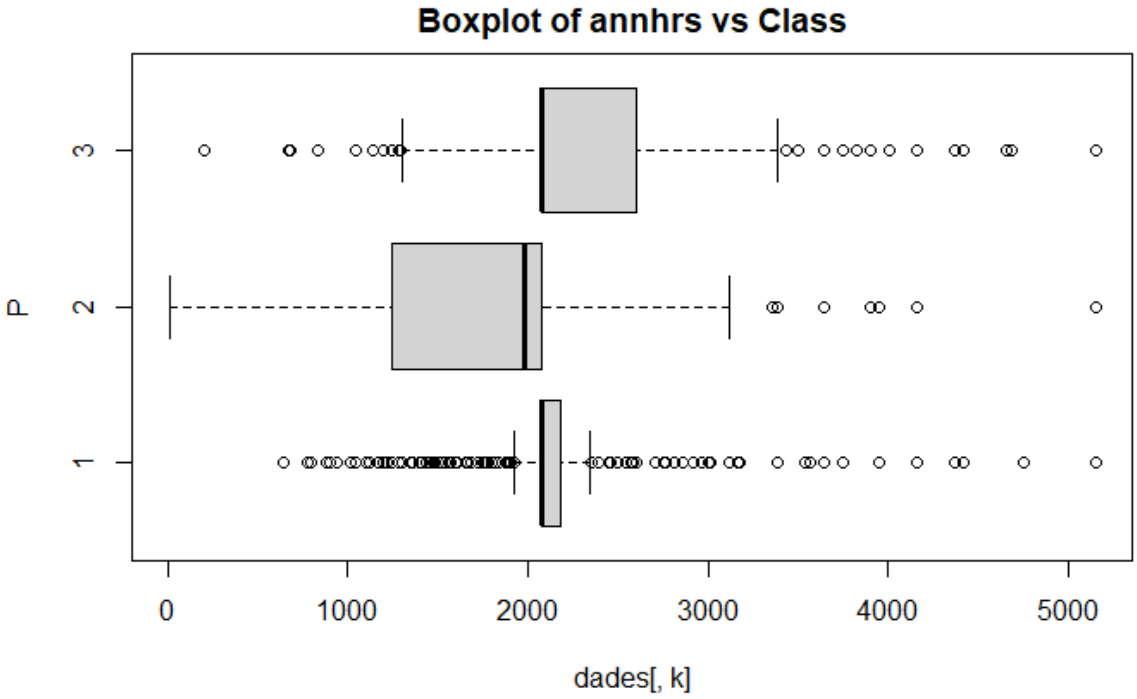


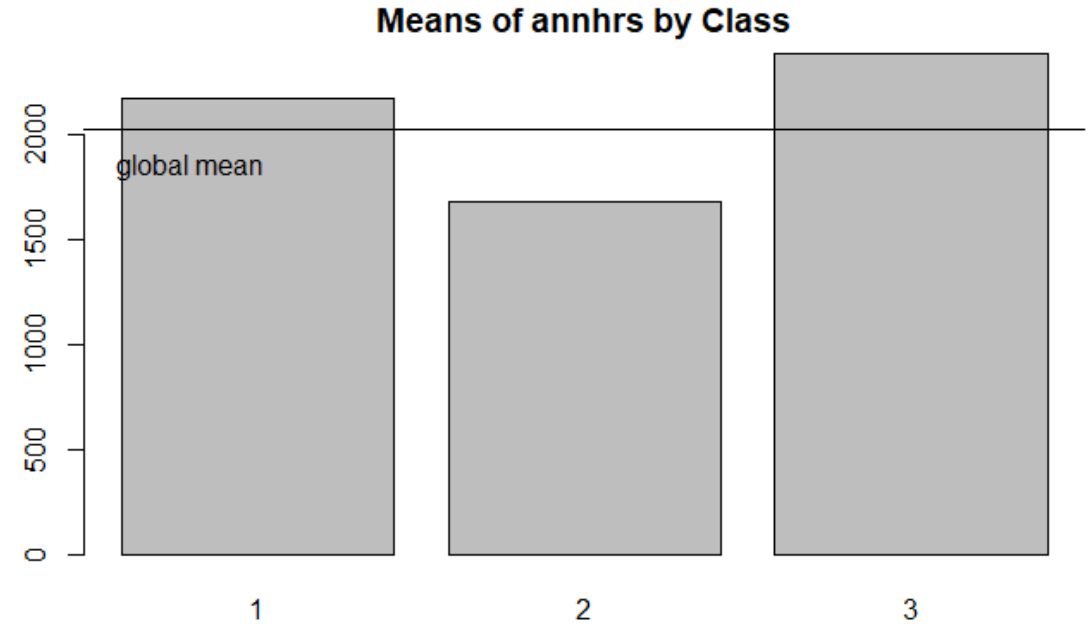


In this case we can easily see there is a big difference between the three clusters. All the outliers belong to the third cluster, which at the same time is the one with the highest mean income wage. Cluster one has a much smaller mean than cluster three, but it's very close to the global mean and it's more than double than the mean of cluster two, which has the lowest mean of them all.

The outliers are clearly the reason why cluster three has such a big difference, but still, there's a lot of information to take away from this variable for characterizing the clusters.

6.1.6 Annhrs

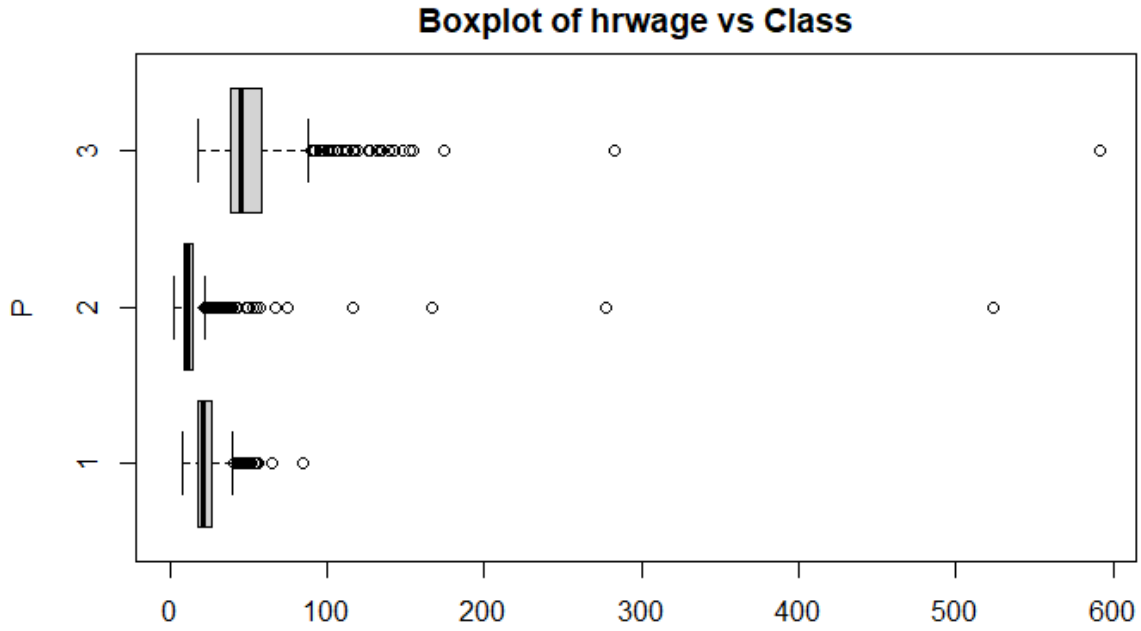


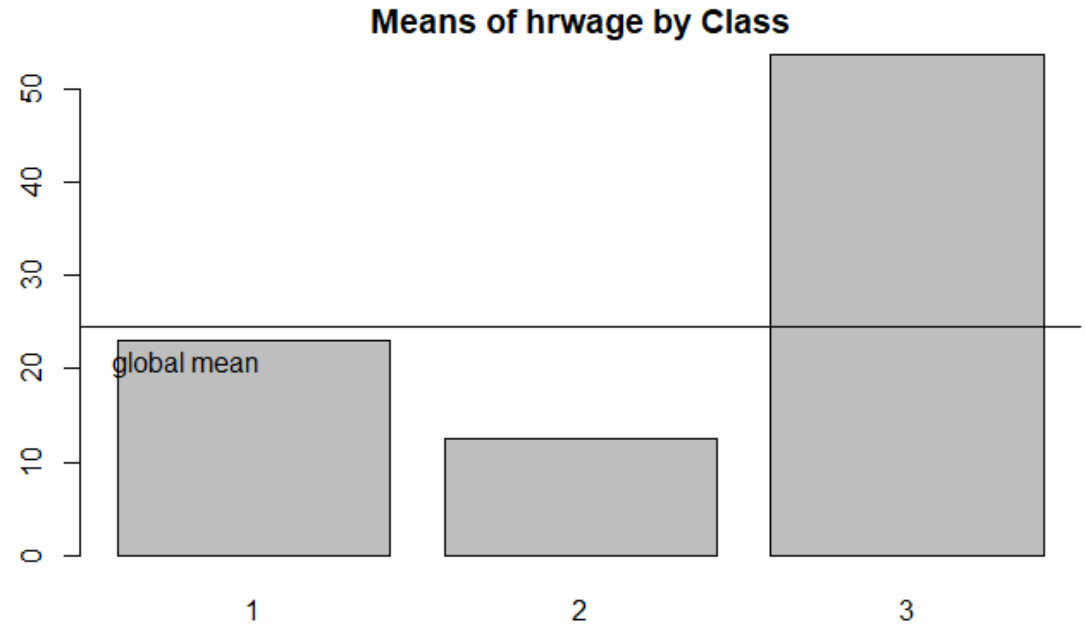


Again, cluster two is the one with the lowest mean bringing the global mean down, and cluster three is slightly higher than cluster one. These plots give similar information to the ones in Uhrswork and Wkswors, which is natural, since with those two we could obtain an accurate estimation of annhrs.

All of the clusters have a considerable number of outliers but cluster one is the one with the most, specially on the side lower than the mean.

6.1.7 Hrwage





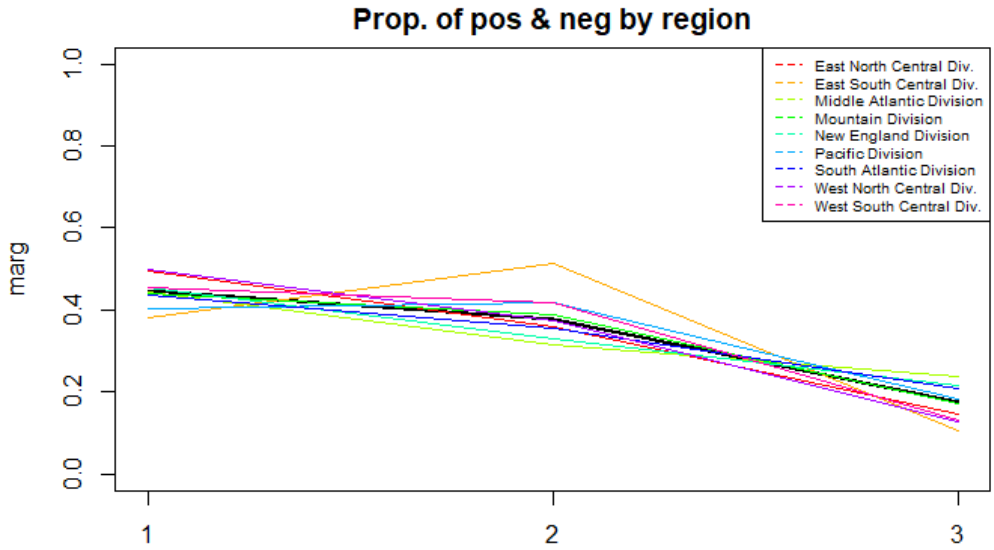
The plot with the means is pretty similar to the one from Incwage, which makes sense. As we can see, again, cluster two is the one with the lowest mean, cluster one is very close to the global mean, and it's double than cluster two. Cluster three is the highest, again, the reason could be the outliers Which all have high values and are probably driving the mean up.

Something interesting about this one compared to Incwage is that cluster 2 does have some really high outliers, which wasn't reflected in Incwage, the reason for that might be that they earn a lot per hour but work less hours annually.

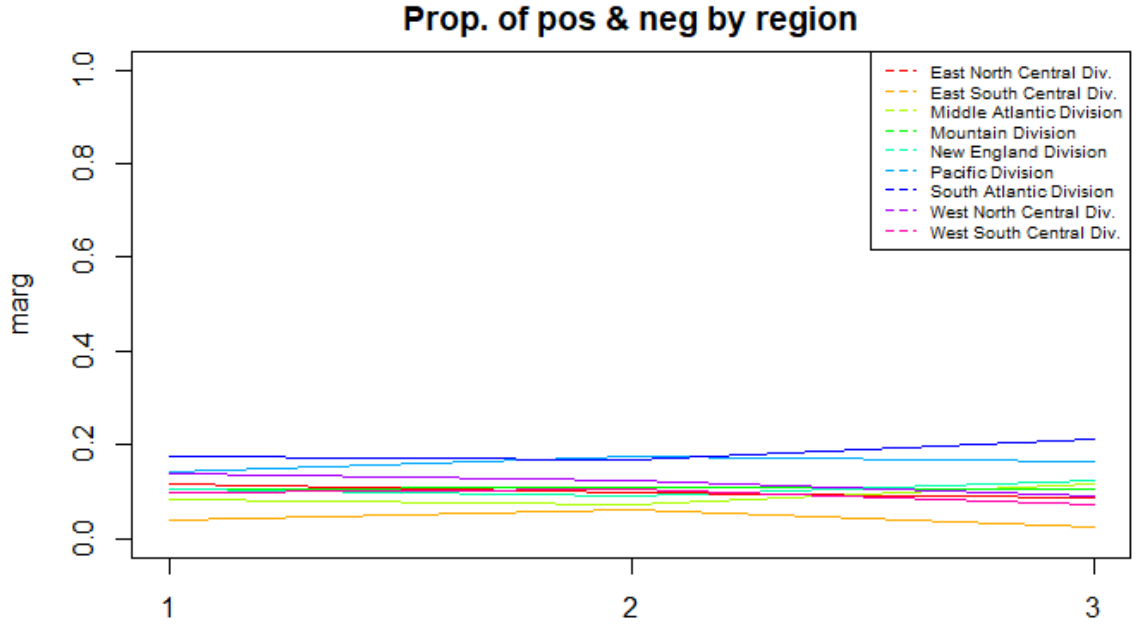
6.2 Categorical variables

6.2.1 Region

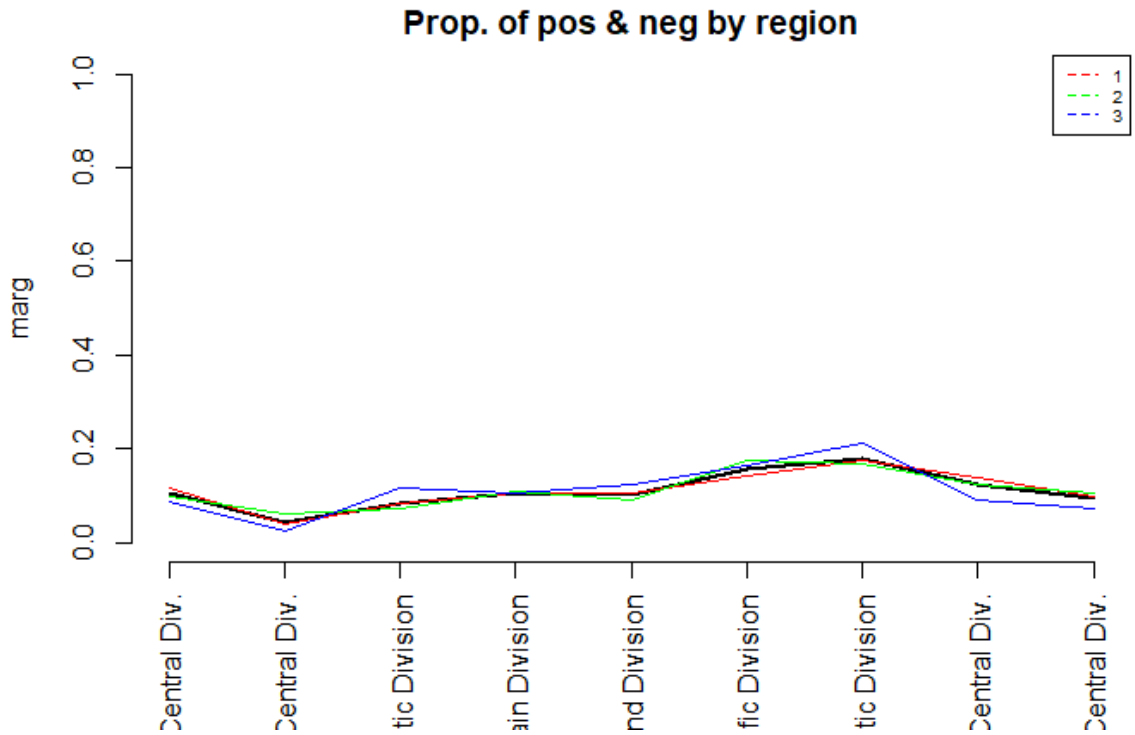
For every categorical variable, we obtained three different snake plots. Before getting our conclusions from the plots, we wanted to make a small commentary on how every plot works.



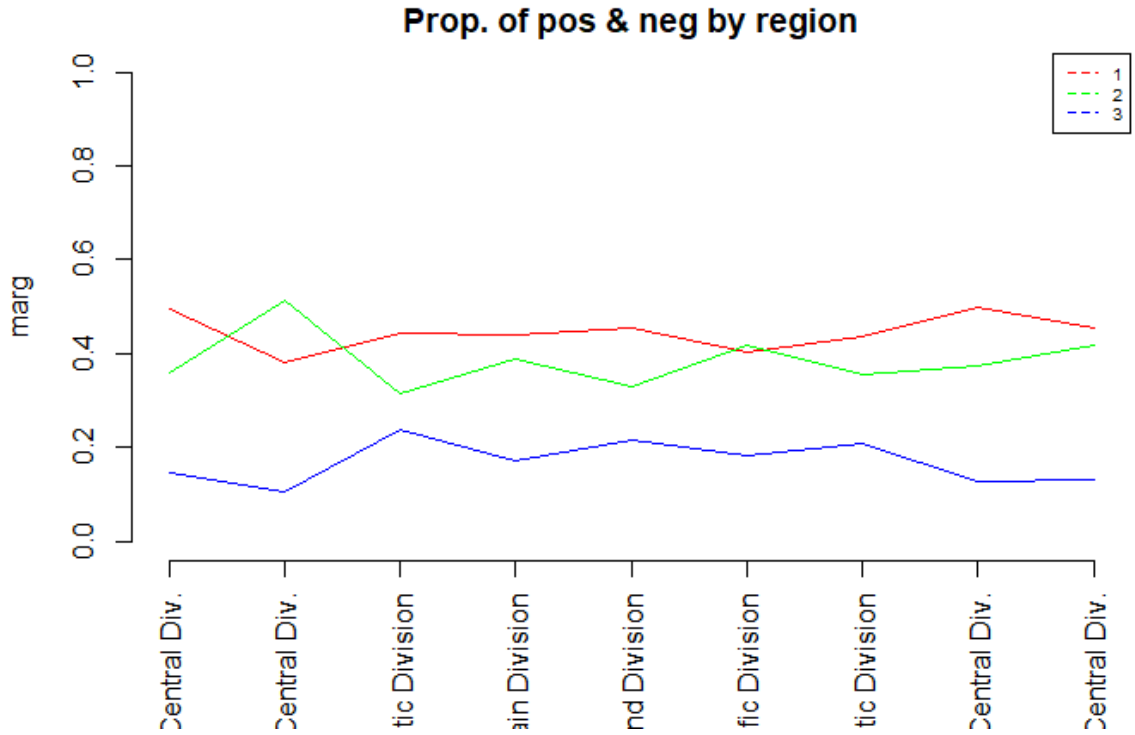
In this case we have a black line which represents the global proportion for the three clusters, for example if the black line was on 0.43 for cluster 1, 0.38 for cluster 2 and 0.18 for cluster three that would mean that from all the individuals 43% are from cluster 1, 38% from cluster 2 and 18% from cluster 3. The colored lines are used to represent the individuals with each value of region, and for every color we can know the percentage that belongs to each cluster.



This graph is different from the first one, here each line represents the percentage of individuals on that cluster that have that label. For example, if the orange line, which represents East South Central Division is at point 0.04, it would mean only 4% of the individuals on the cluster have that value.

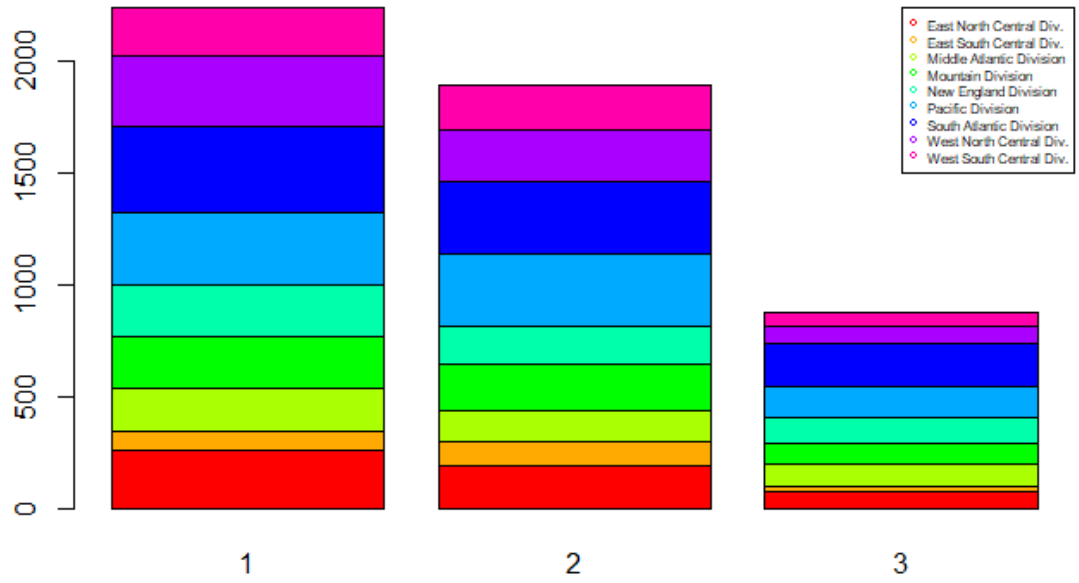


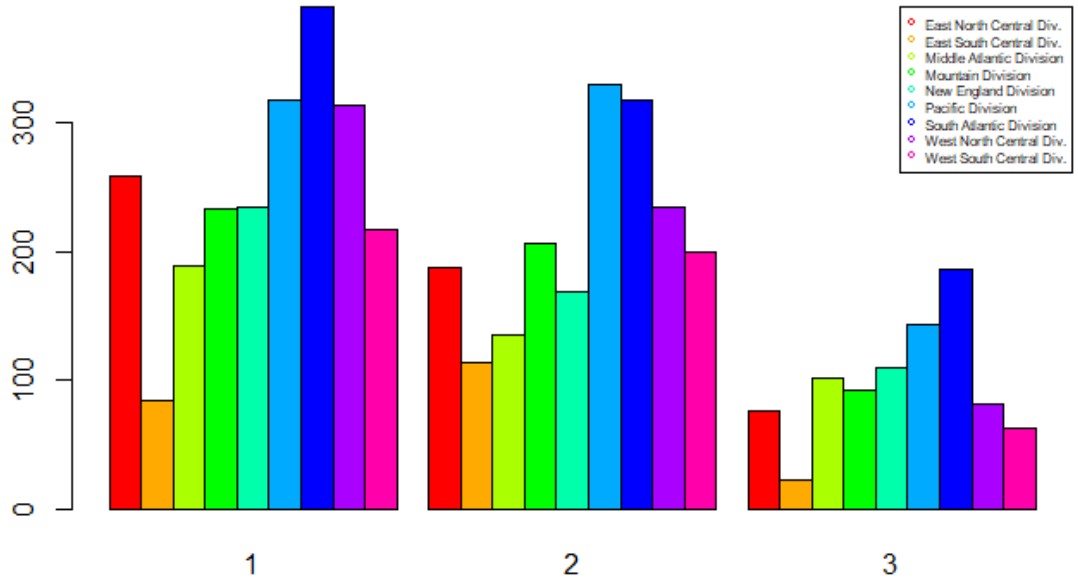
This graph is the same kind as the first one we saw, but it represents a different thing, since in this case the ones on the X axis are the different values the variable can take.



This is the same kind as the second, but again, it represents a different thing, since in this case the ones on the X axis are the different values the variable can take.

Lastly we might also use barplots we obtained, these are pretty self explanatory so we won't go deeper.

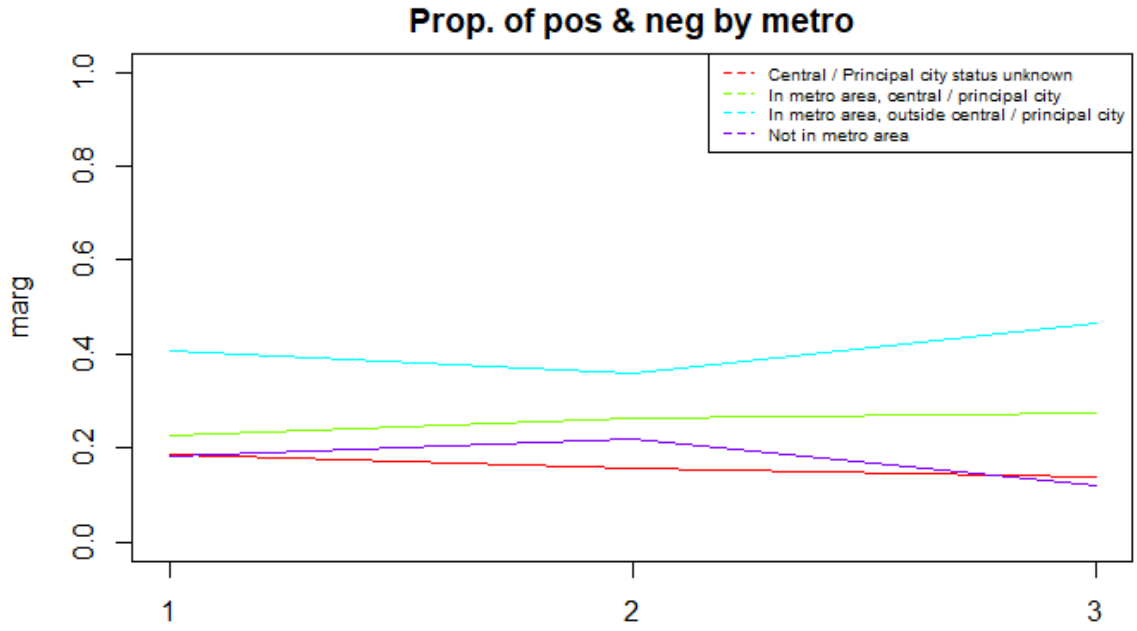


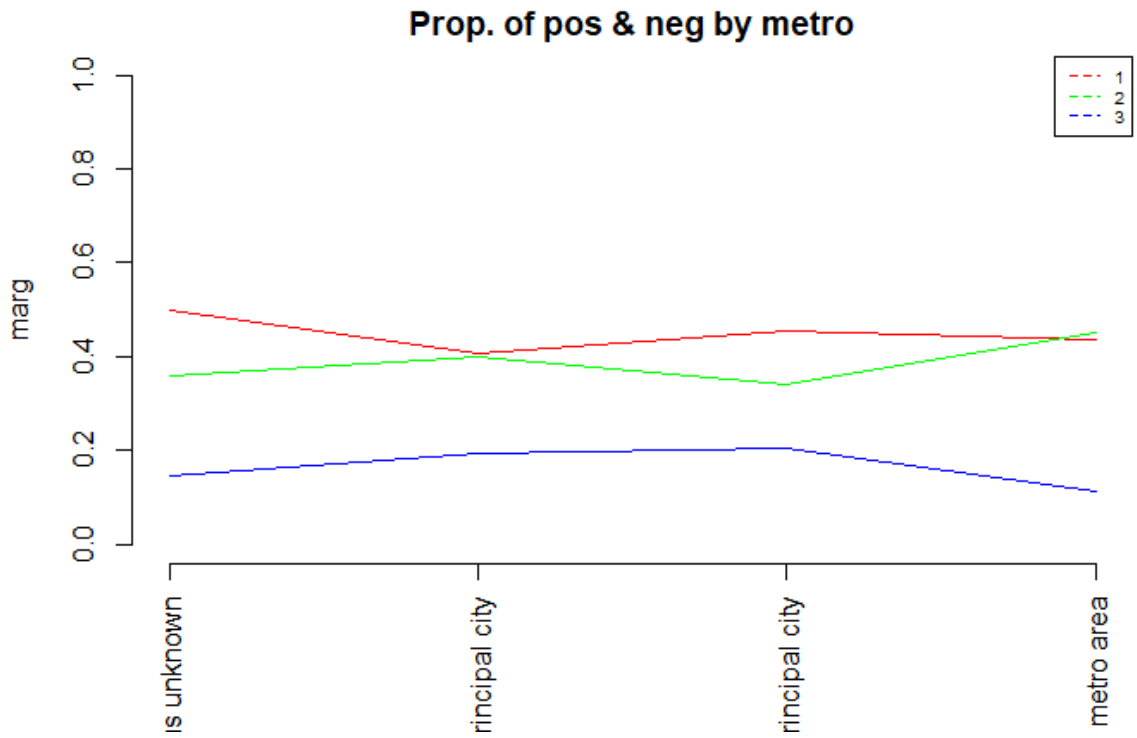


Although we generated all these plots for every variable, we are only going to include some for the analysis, as to not have five different graphs for each variable, with the exception of this first variable, where we wanted to explain every kind of plot.

For regions we can see there isn't a big difference between the different clusters. In clusters one and three the most predominant one is South Atlantic Division, for cluster two that's actually the second one, with the most predominant one being New England Division. For all of them the least common one is East South Central Division, although proportionally cluster two has a higher amount of it.

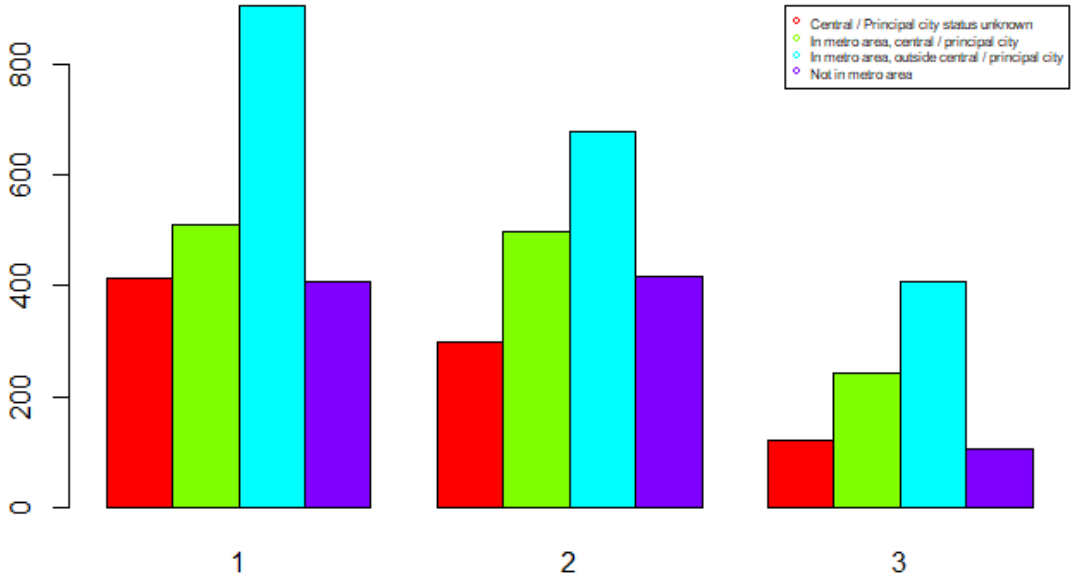
6.2.2 Metro





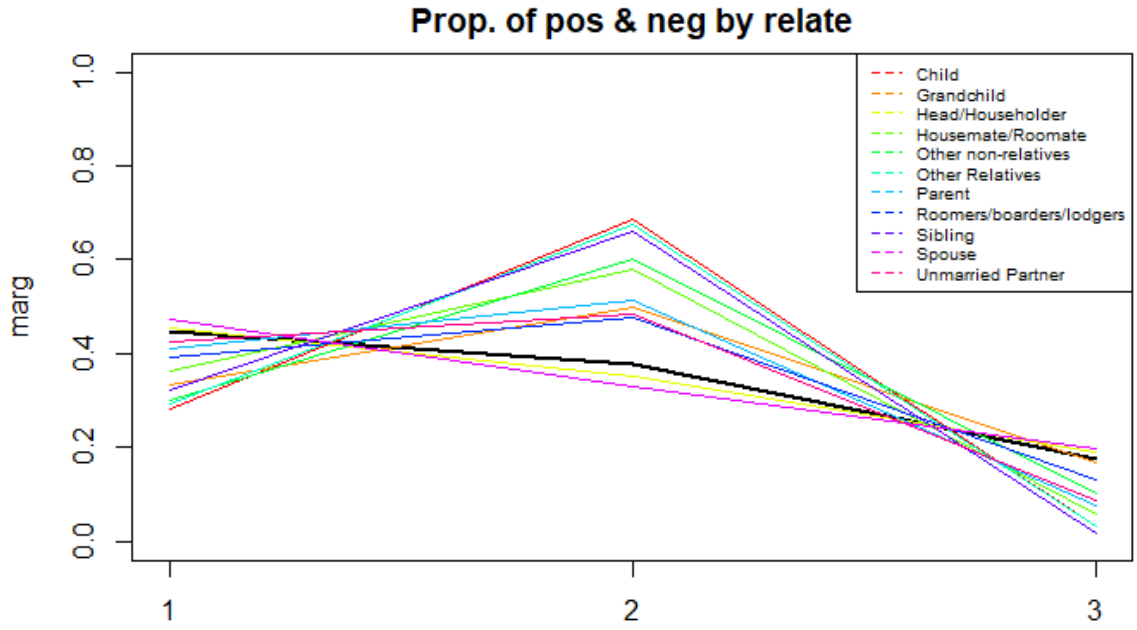
The first plot allows us to see that cluster three has a higher percentage of people who live in metropolitan areas than cluster one and two. Cluster two is the one with the highest amount of people living outside of metropolitan areas.

The differences between the clusters are not too big, so it's not something very defining of each cluster.



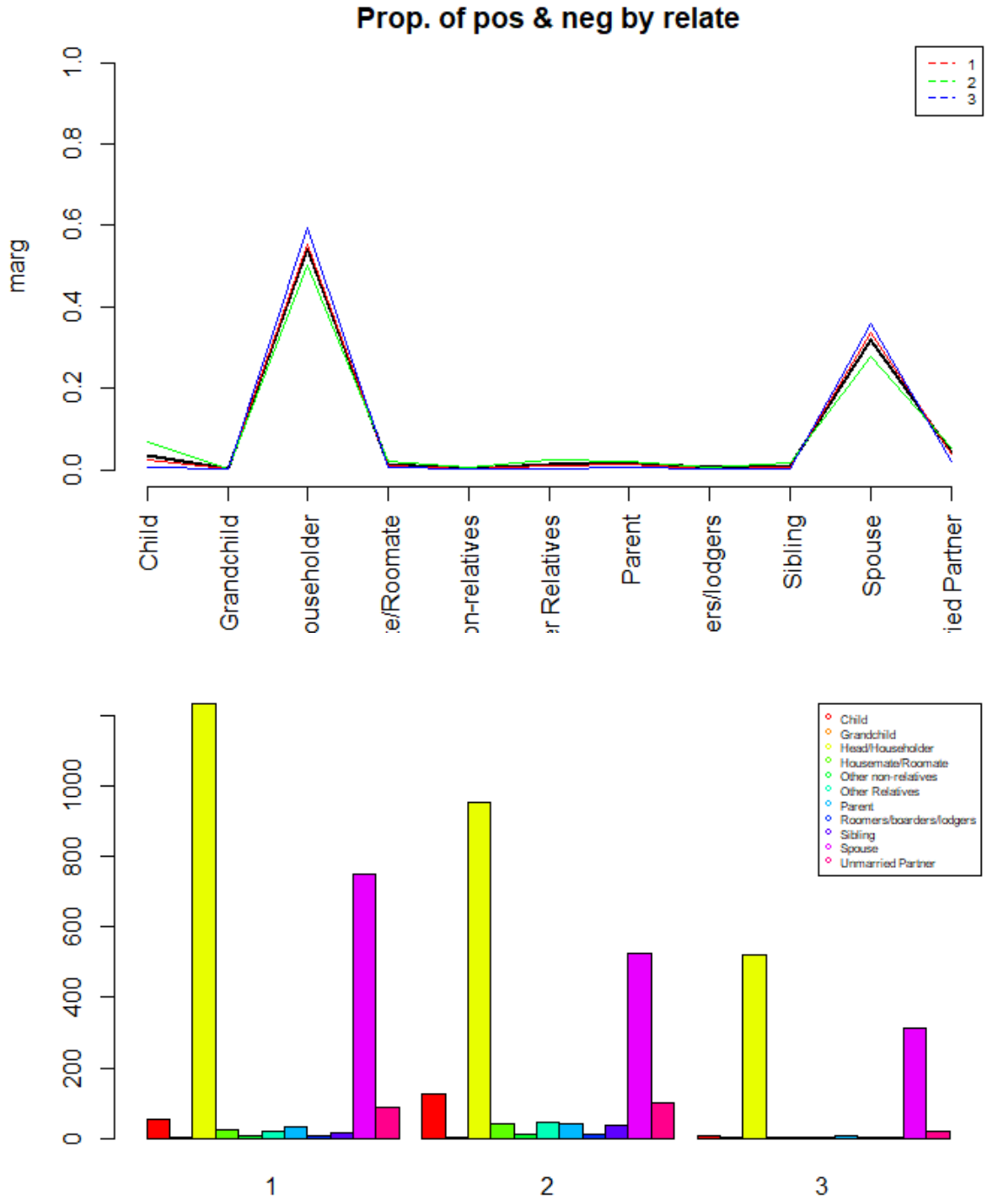
This graph further shows that they are all very similar to each other, and the biggest difference is that cluster three has the lowest number of people not living in a metropolitan area.

6.2.3 Relate



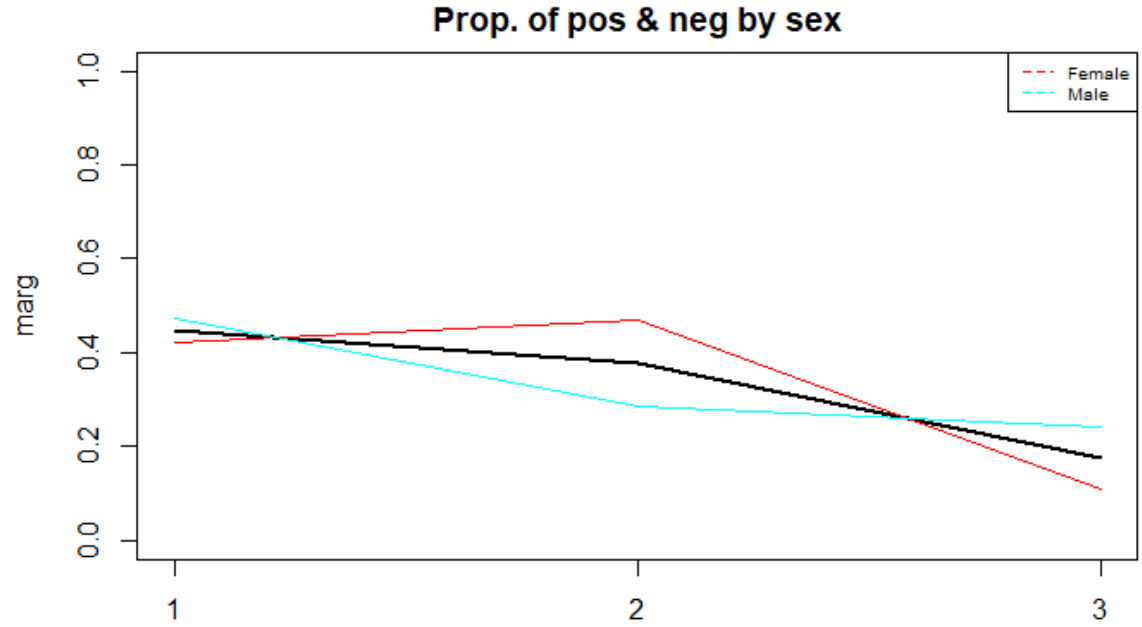
In the first plot we can see that cluster two is the one that presents the biggest differences. Something notable is that they have a big percentage of the individuals whose relationship to the head of the household is “Child”, “Sibling” and “Grandchild”. Those would make us

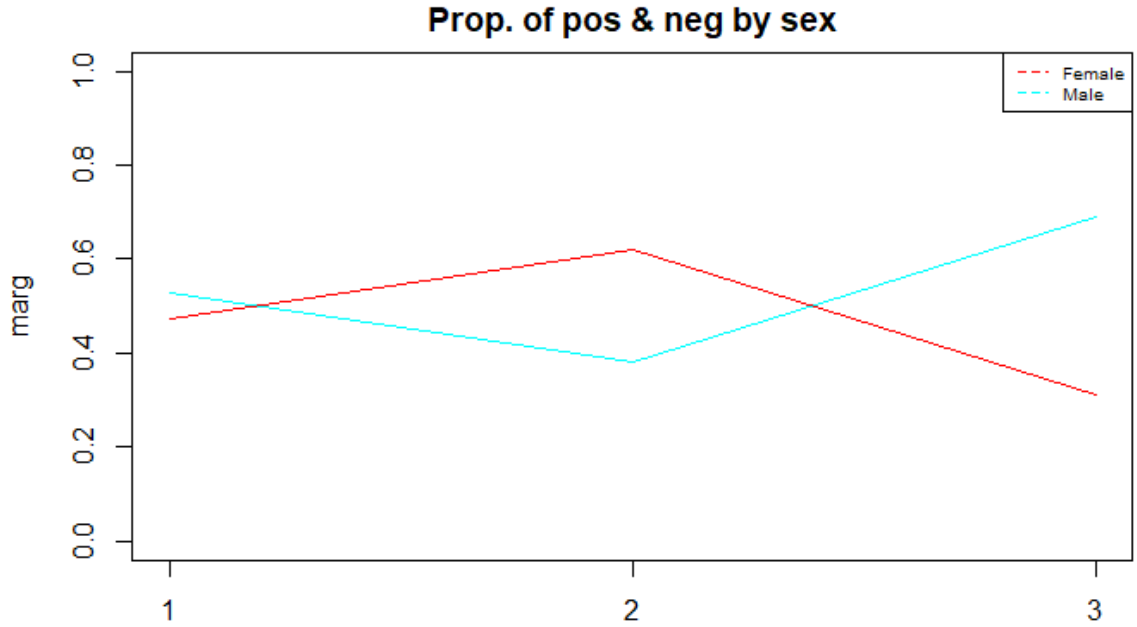
think about younger people or people who are not the main source of income of the house.

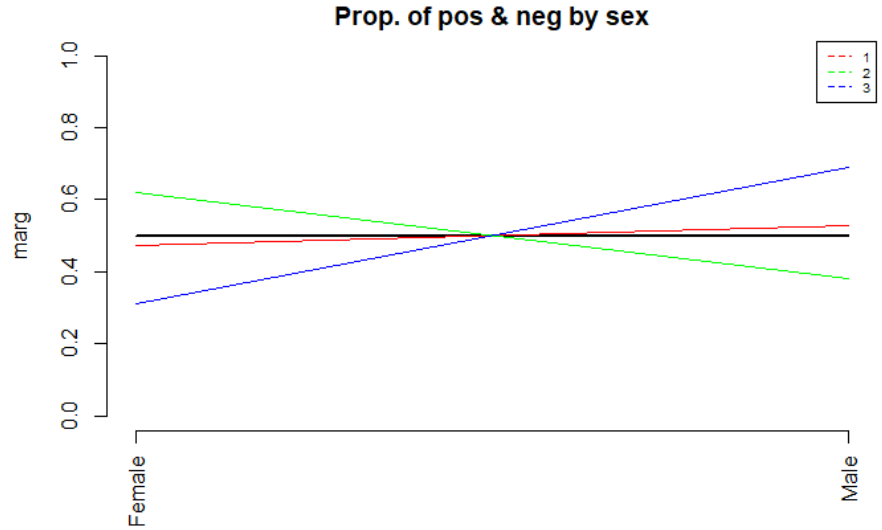


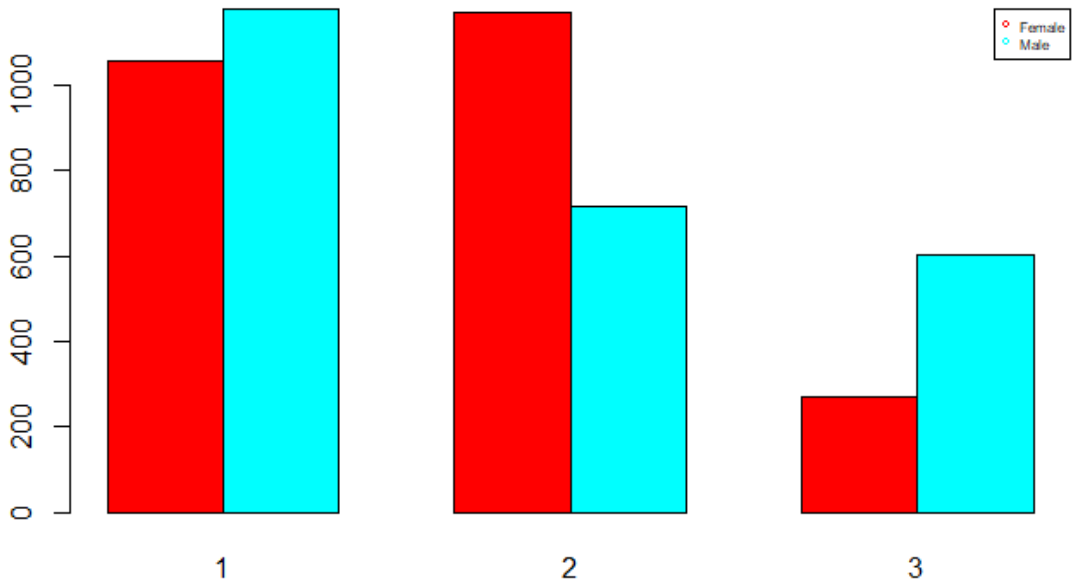
In the second plot we see that actually those differences are not so important because in the end the ones that represent the biggest number of individuals are the values “Head/ Householder/” and “Spouse”.

6.2.4 Sex


Prop. of pos & neg by sex
Female
Male
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3


Prop. of pos & neg by sex
Female
Male
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3


Prop. of pos & neg by sex
1
2
3
marg
0.0
0.2
0.4
0.6
0.8
1.0
Female
Male



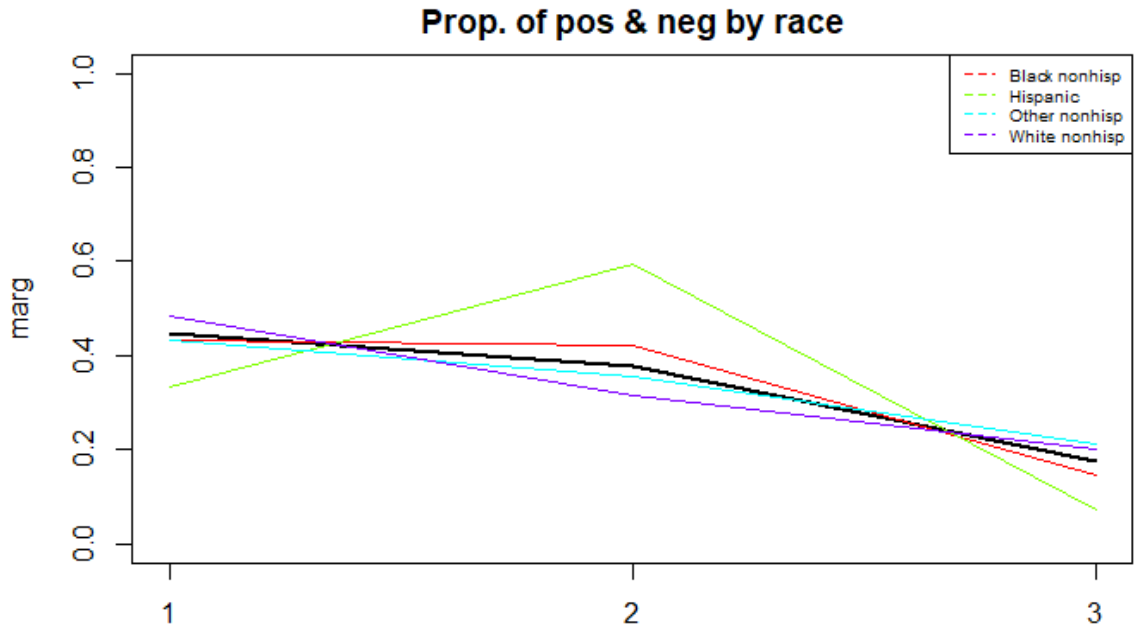
As we can see in every single plot, cluster three have a significantly higher percentage of men among their individuals than the rest of clusters, with almost 70% of their individuals being males. They have more than double males than females.

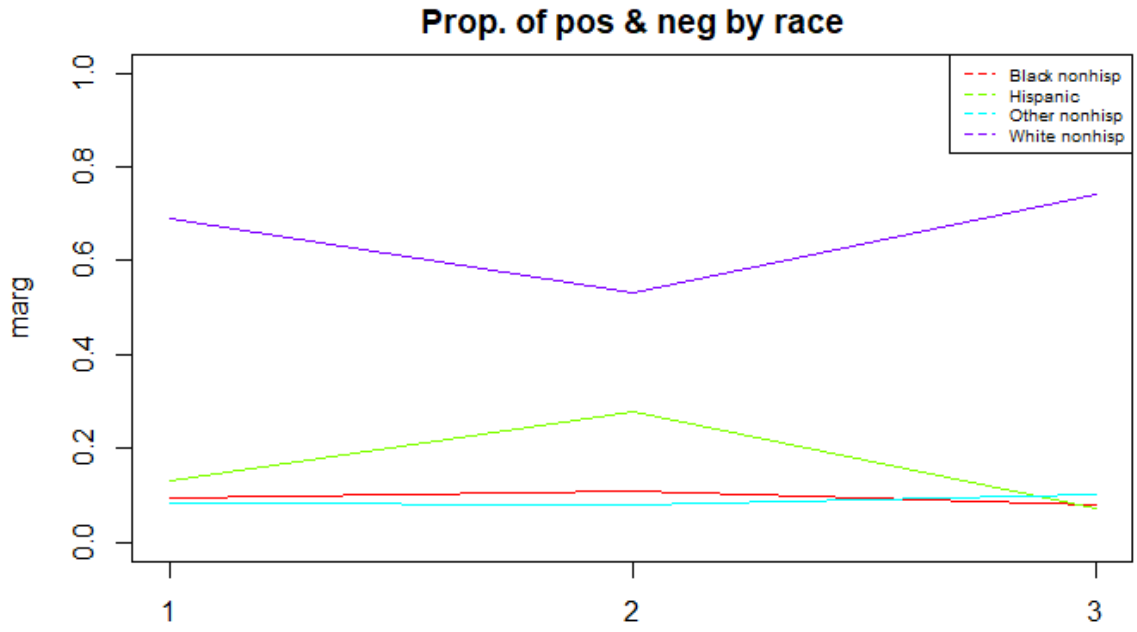
Cluster two has a significantly higher number of females than males, with around 60% females and 40% males.

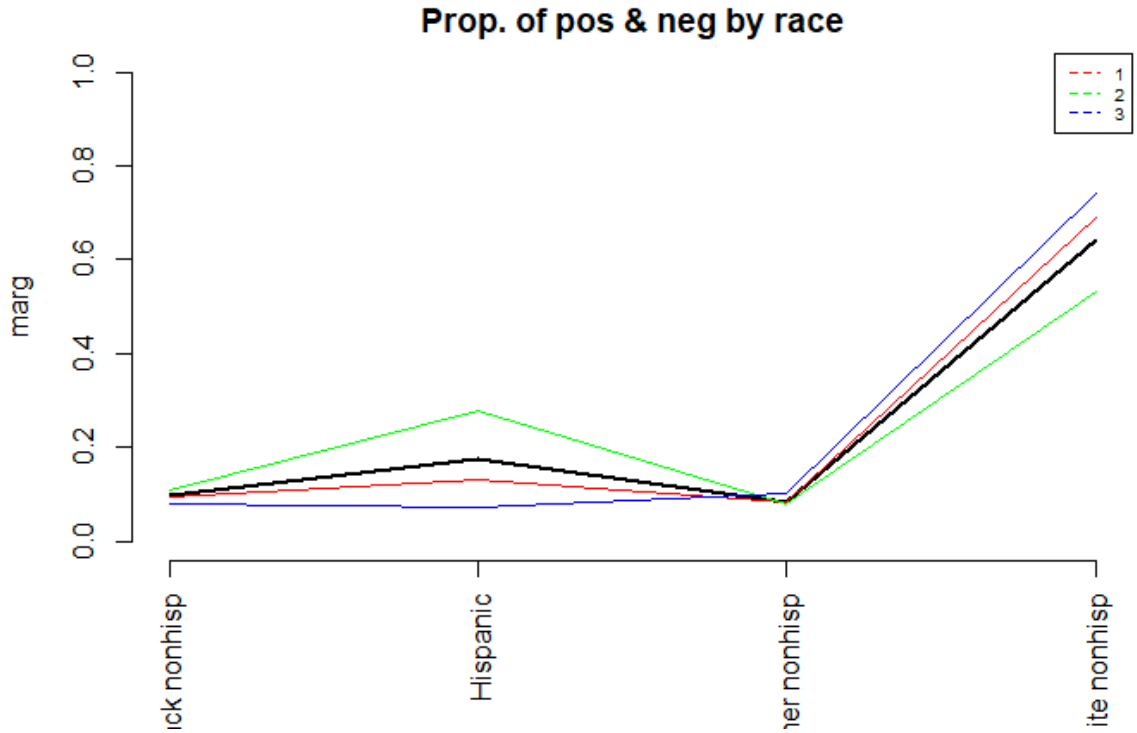
Cluster one is the most balanced one, but it does have slightly more males than females.

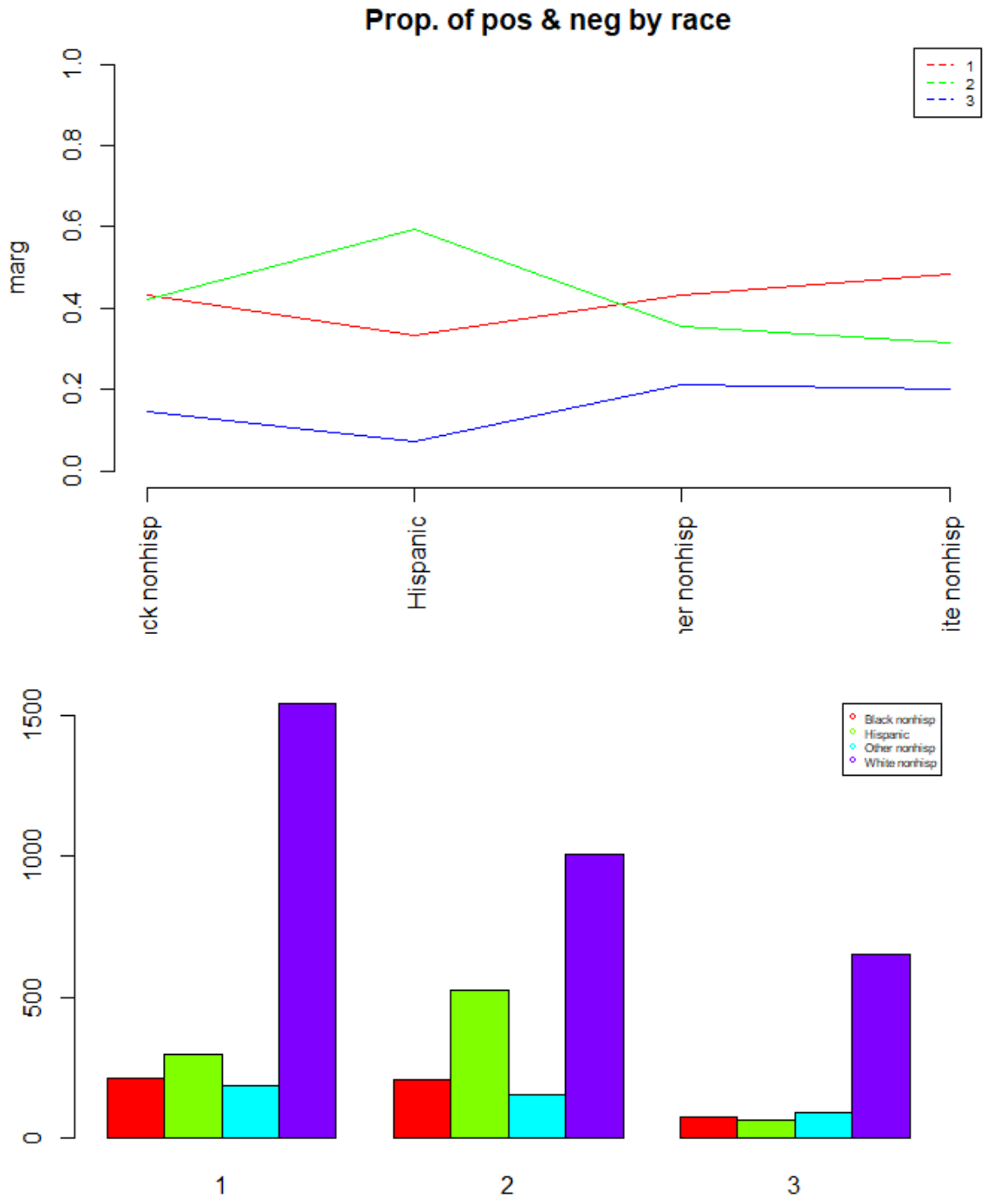
These findings are very relevant for our topic.

6.2.5 Race




Prop. of pos & neg by race
Black nonhisp
Hispanic
Other nonhisp
White nonhisp
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3

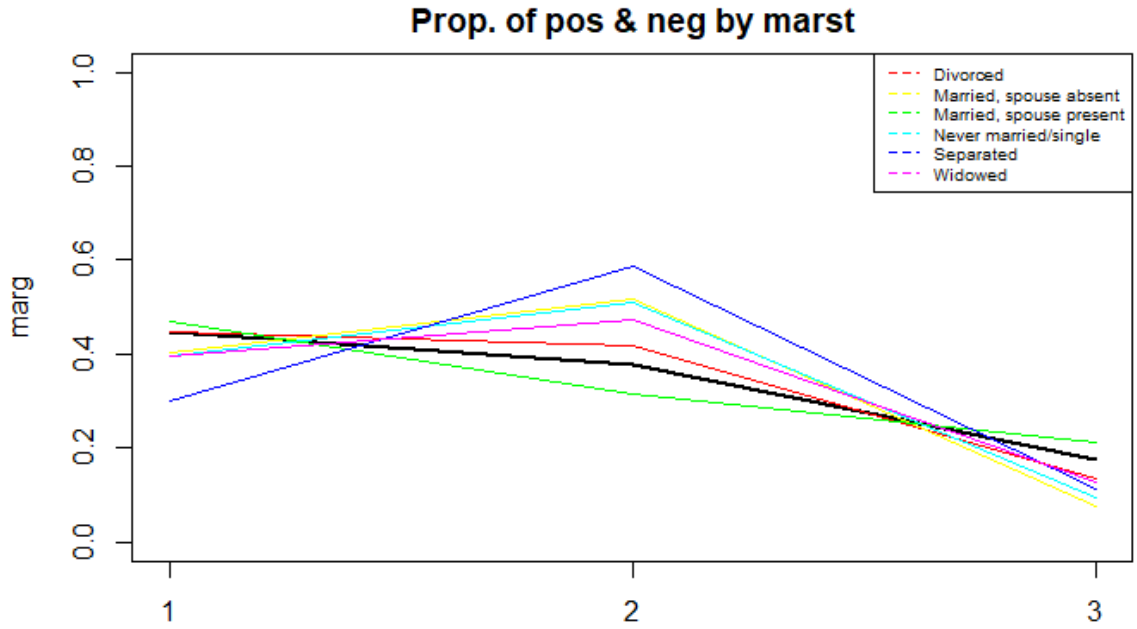

Prop. of pos & neg by race
1
2
3
marg
0.0
0.2
0.4
0.6
0.8
1.0
ıck nonhisp
Hispanic
ıer nonhisp
ite nonhisp

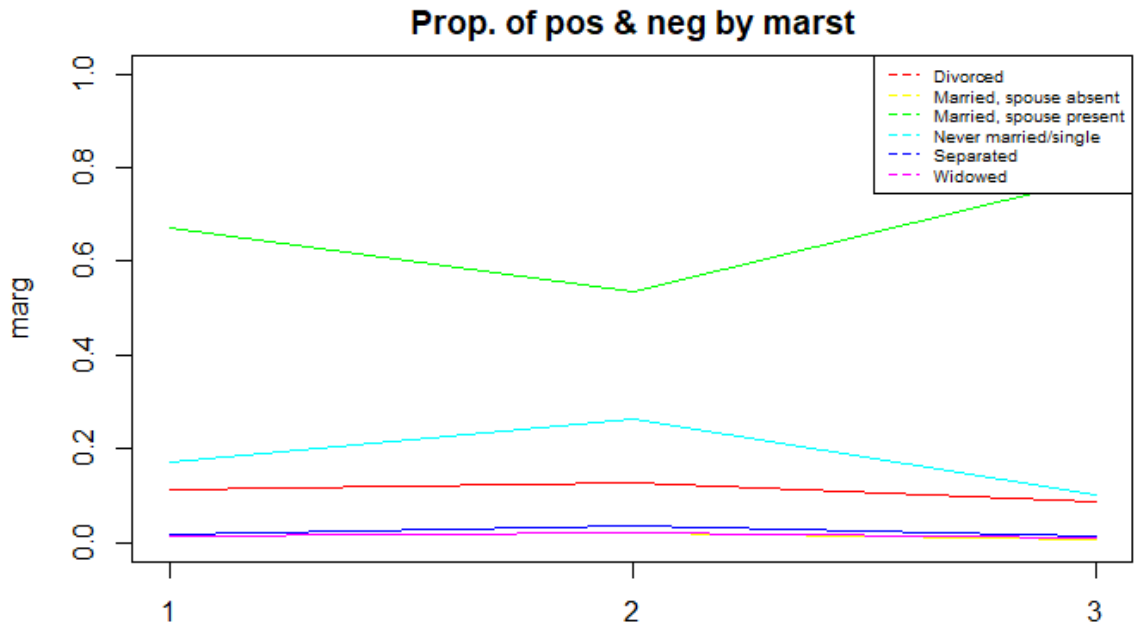


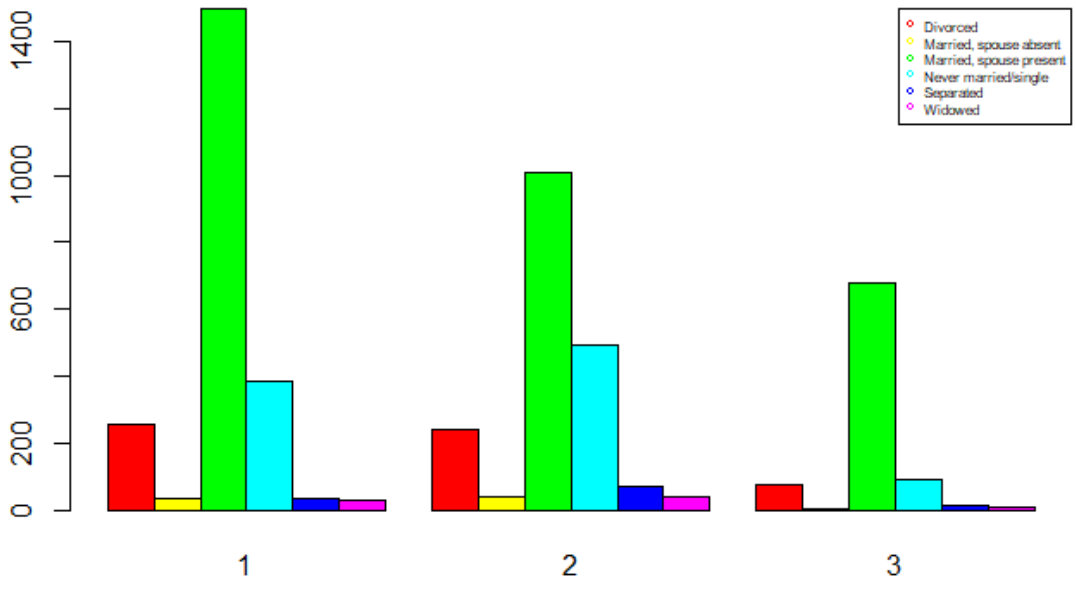
There are also two significant differences which can be observed from the race variable. Cluster two has the highest number of hispanic individuals and proportionally the lowest percentage of white non hispanic individuals.

The rest are pretty much the same for all the clusters, it might be interesting to know that cluster 3 has the lowest percentage of hispanic individuals.

6.2.6 Marst


Prop. of pos & neg by marst
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
Divorced
Married, spouse absent
Married, spouse present
Never married/single
Separated
Widowed

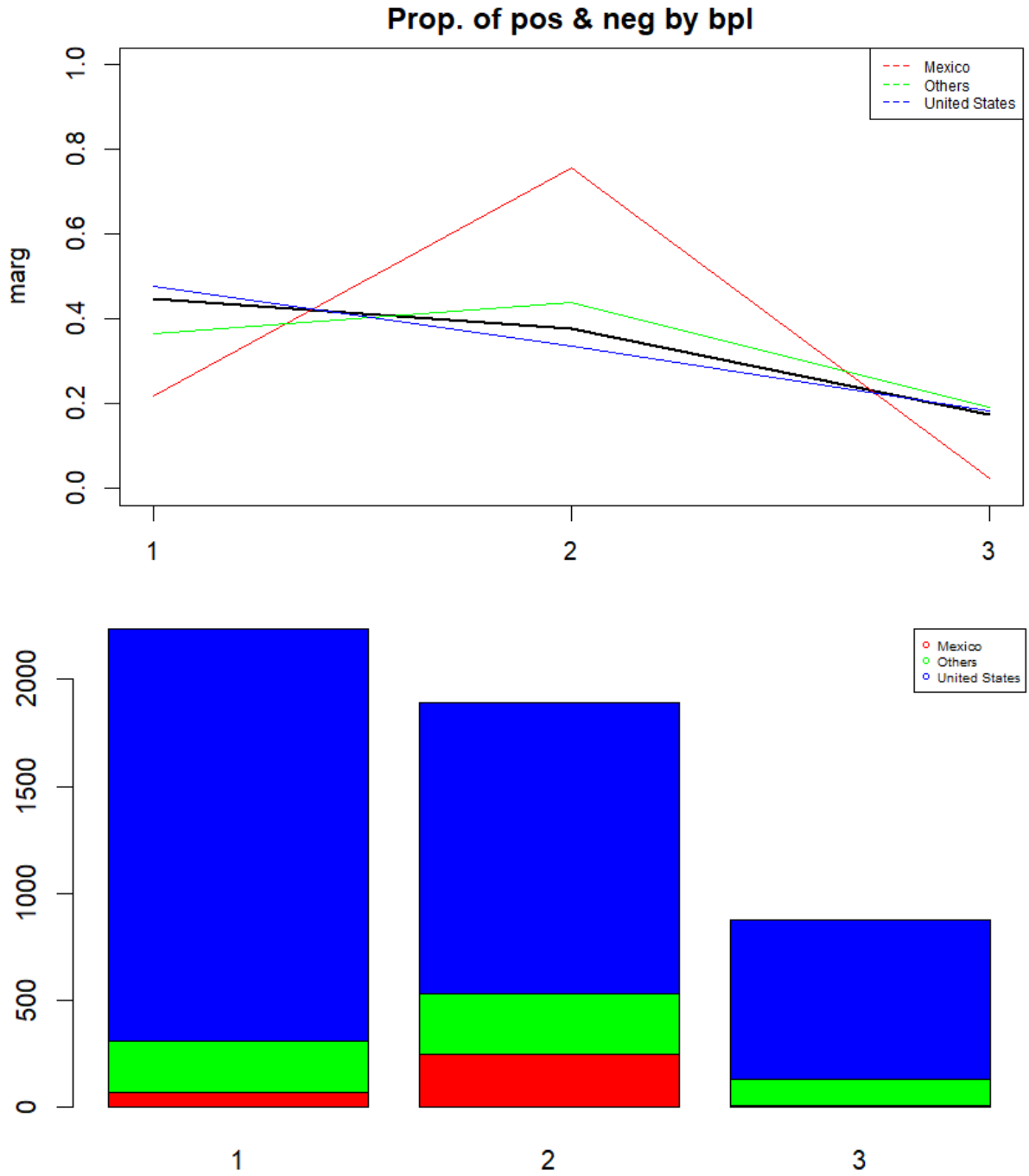

Prop. of pos & neg by marst
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
Divorced
Married, spouse absent
Married, spouse present
Never married/single
Separated
Widowed


0
200
600
1000
1400
1
2
3
Divorced
Married, spouse absent
Married, spouse present
Never married/single
Separated
Widowed

Cluster two has the highest percentage of divorced, never married, separated, widowed, and married with absent spouse individuals and the lowest percentage of married individuals with present spouse.

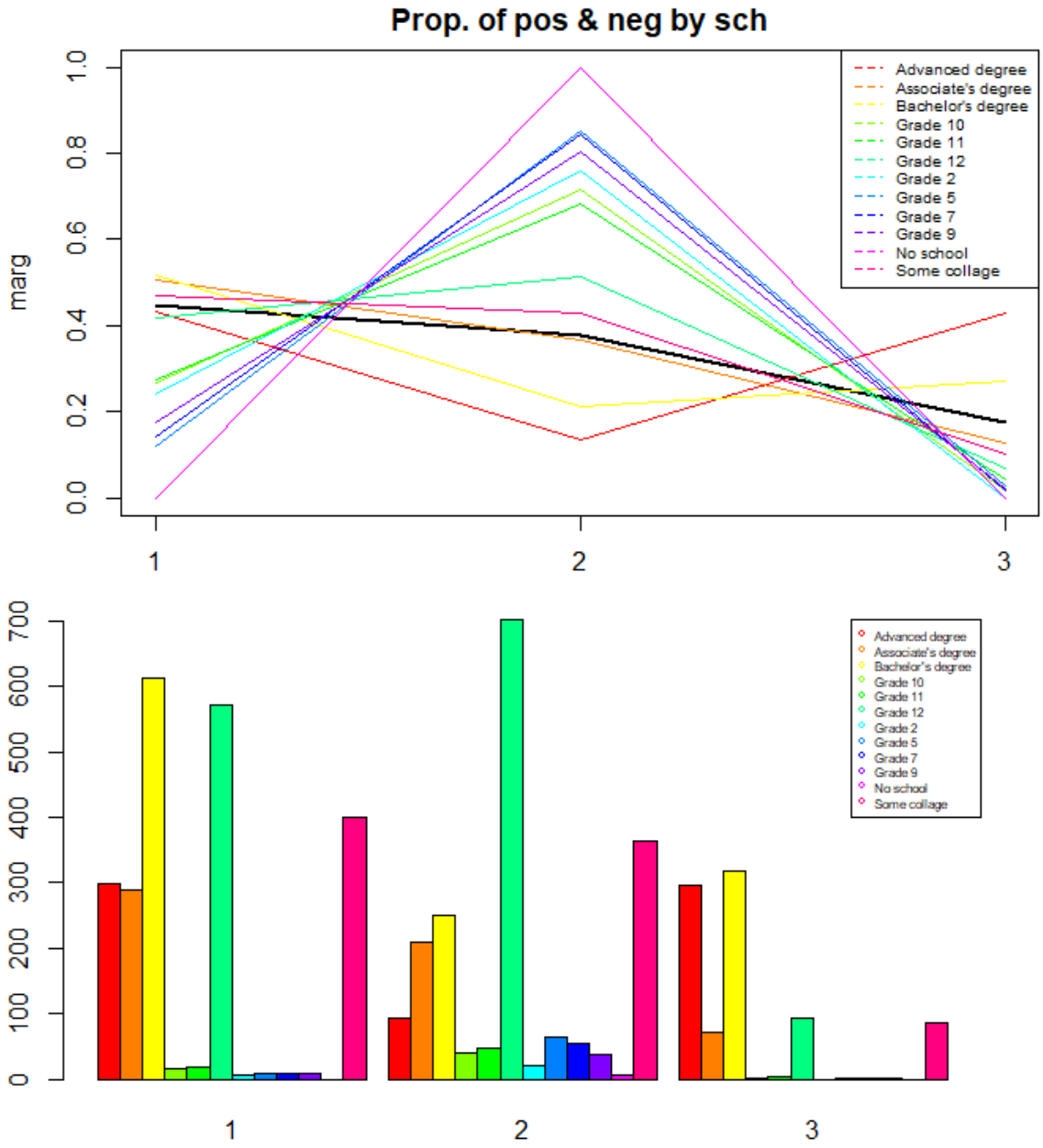
Cluster three has the lowest percentage of never married / single as well as married with absent spouses.

6.2.7 Bpl



Cluster two has the highest number of individuals born in Mexico, while cluster three has the lowest number of them, with almost no individuals born there. Cluster two is also the one with the most individuals born outside of Mexico and the United States. Cluster one is actually the one with the lowest percentage of individuals born in places other than Mexico and United States

6.2.8 Sch

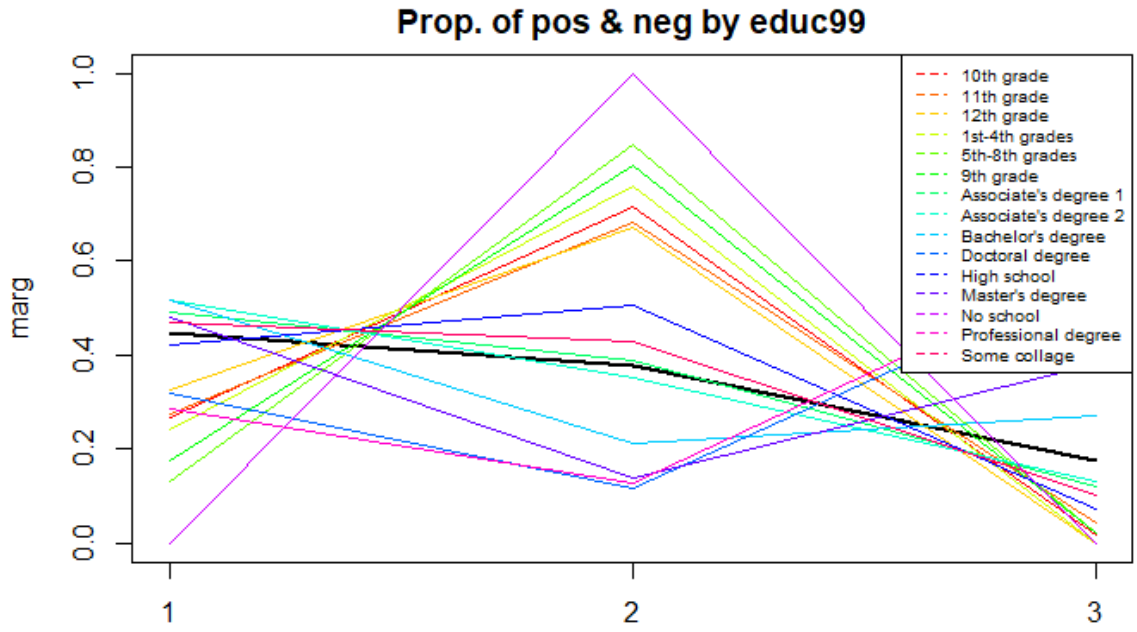


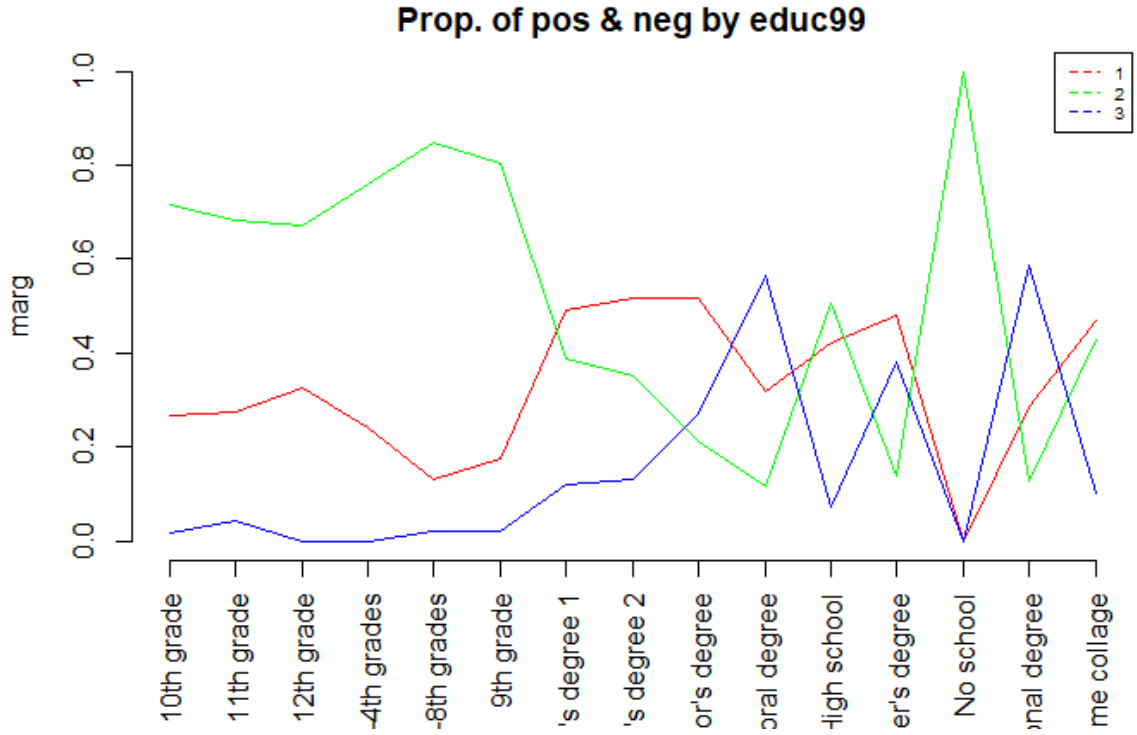
Cluster three has the lowest number of individuals that decided to not pursue further education after high school. It also has the highest percentage of individuals with an advanced or bachelor's degree.

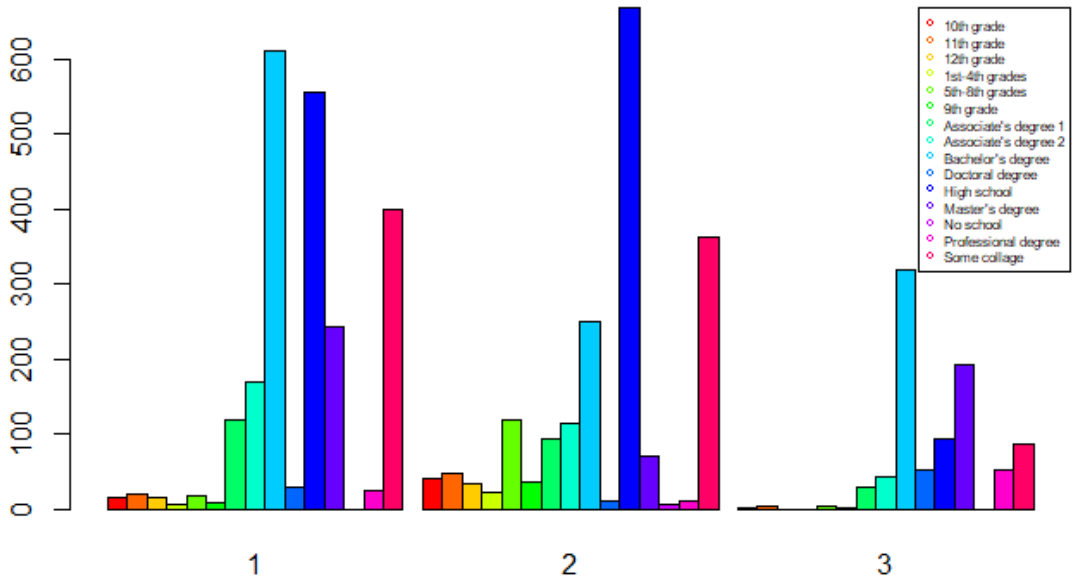
Cluster two has the highest number of individuals that decided to finish their education before or right after graduating high school. They also have most of the individuals who never attended school, however this group represents a really small number of individuals.

Cluster one has a pretty high percentage of individuals that finished their education after highschool, but still, they also have a big amount of individuals that pursued some further education.

6.2.9 Educ99

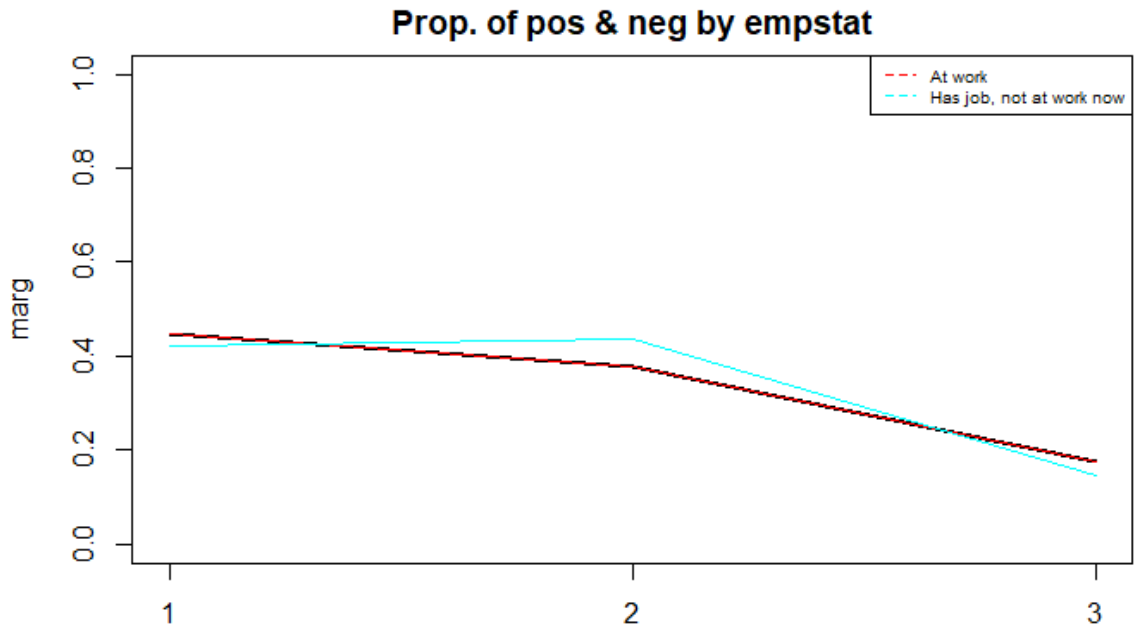

Prop. of pos & neg by educ99
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
10th grade
11th grade
12th grade
1st-4th grades
5th-8th grades
9th grade
Associate's degree 1
Associate's degree 2
Bachelor's degree
Doctoral degree
High school
Master's degree
No school
Professional degree
Some collage

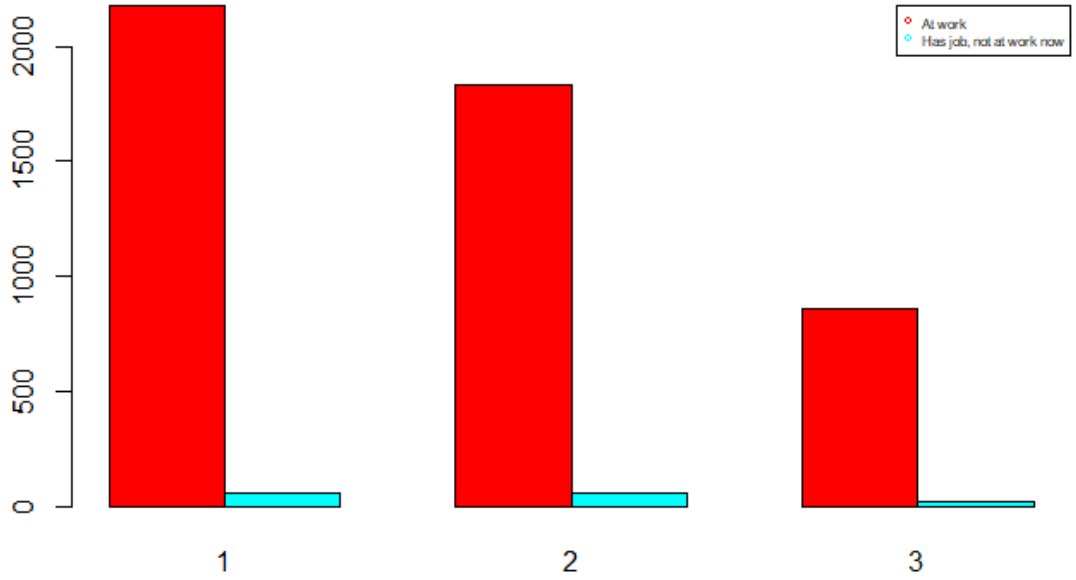

Prop. of pos & neg by educ99
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
10th grade
11th grade
12th grade
-4th grades
-8th grades
9th grade
's degree 1
's degree 2
or's degree
oral degree
High school
er's degree
No school
nal degree
me collage



The information from this variable is very similar to the one we got from Sch. There are some slight additions we know that cluster three is the cluster with the highest number of individuals with a doctoral, and professional degree, and also the one with the highest percentage of individuals with a master's degree.

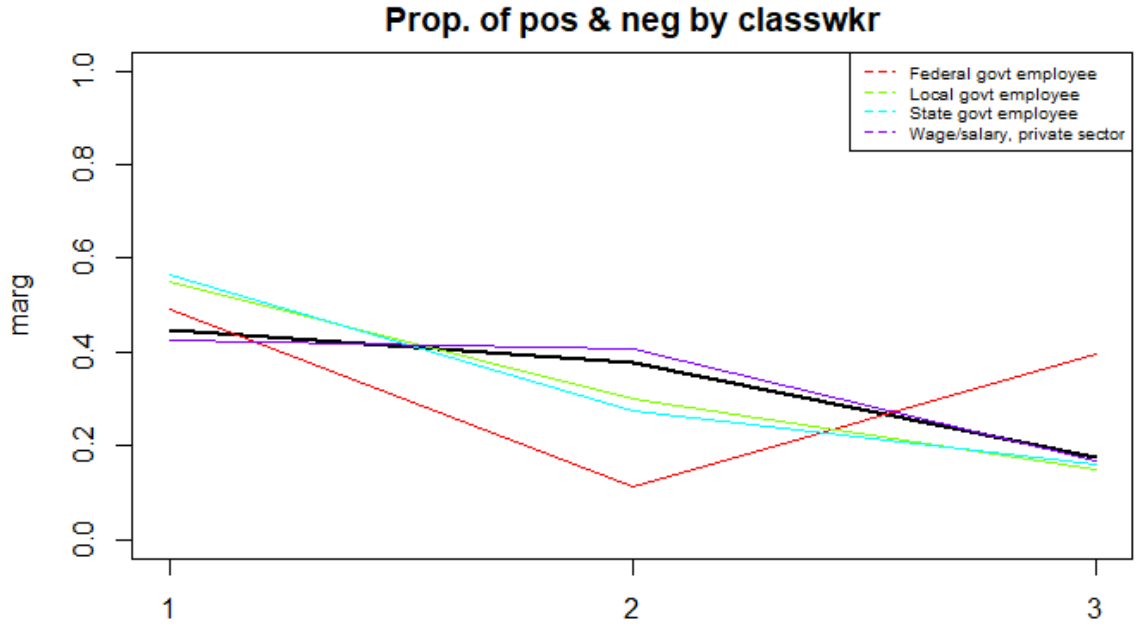
6.2.10 Empstat

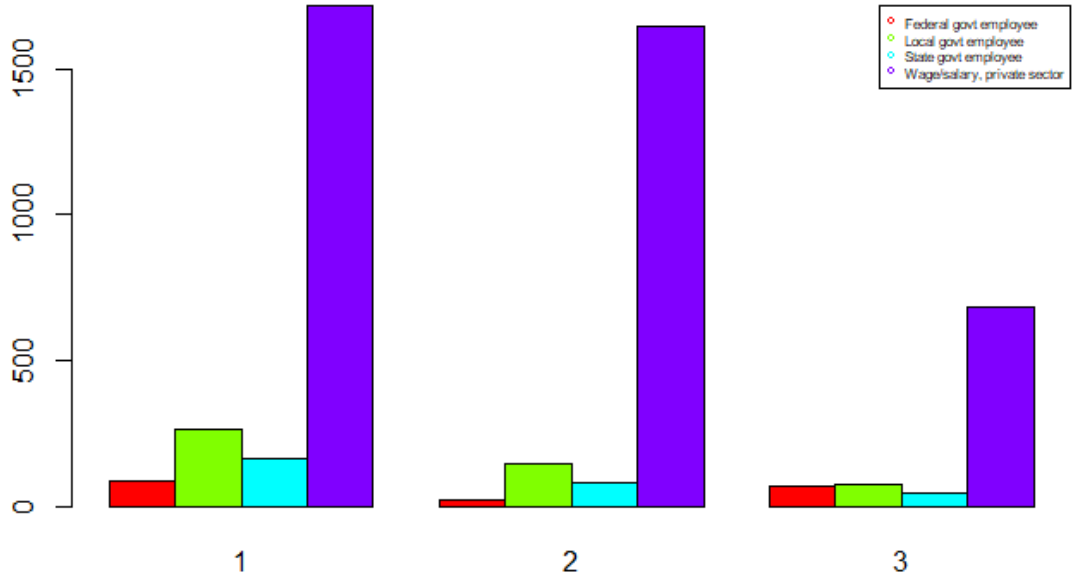




The three clusters have very similar values for this value, cluster two has a slightly higher proportion of people with a job but not currently working, but the difference is not significant.

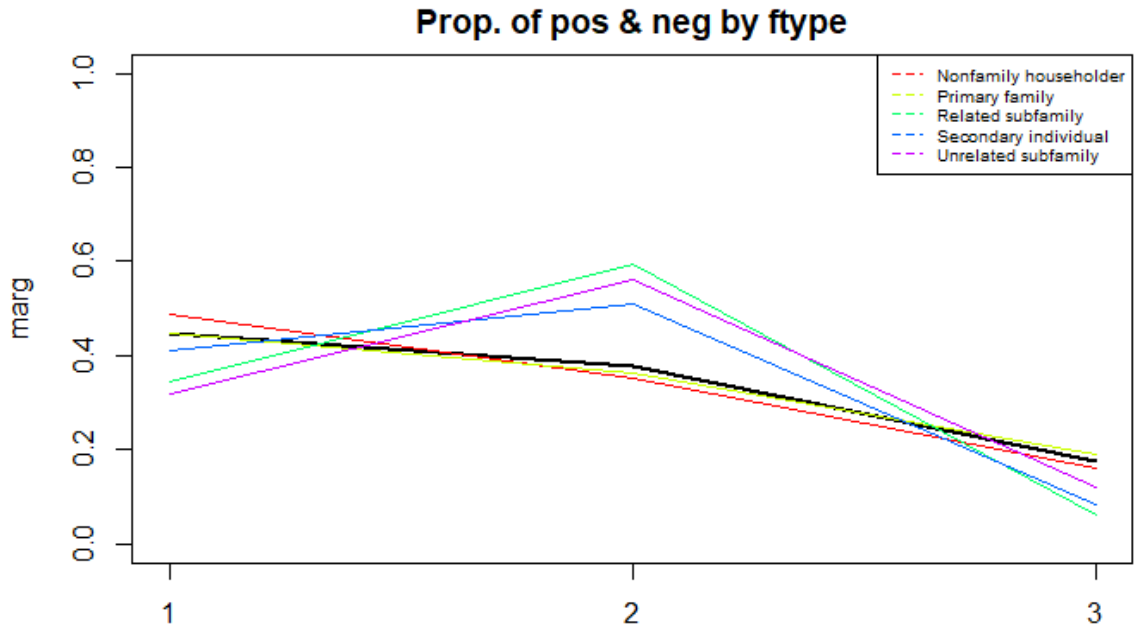
6.2.11 Classwkr

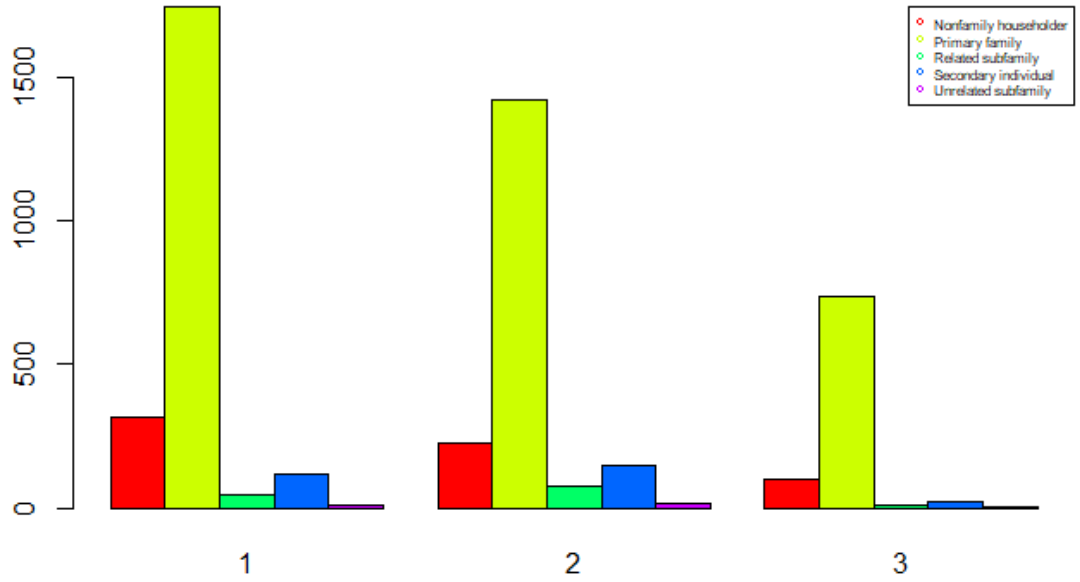




Wage / salary, private sector is the most frequent value among the three clusters. There is not a big difference between each cluster's value for this variable, but still we can see cluster two has the lowest percentage of individuals who are federal government employees, while cluster three has the highest one.

6.2.12 Ftype



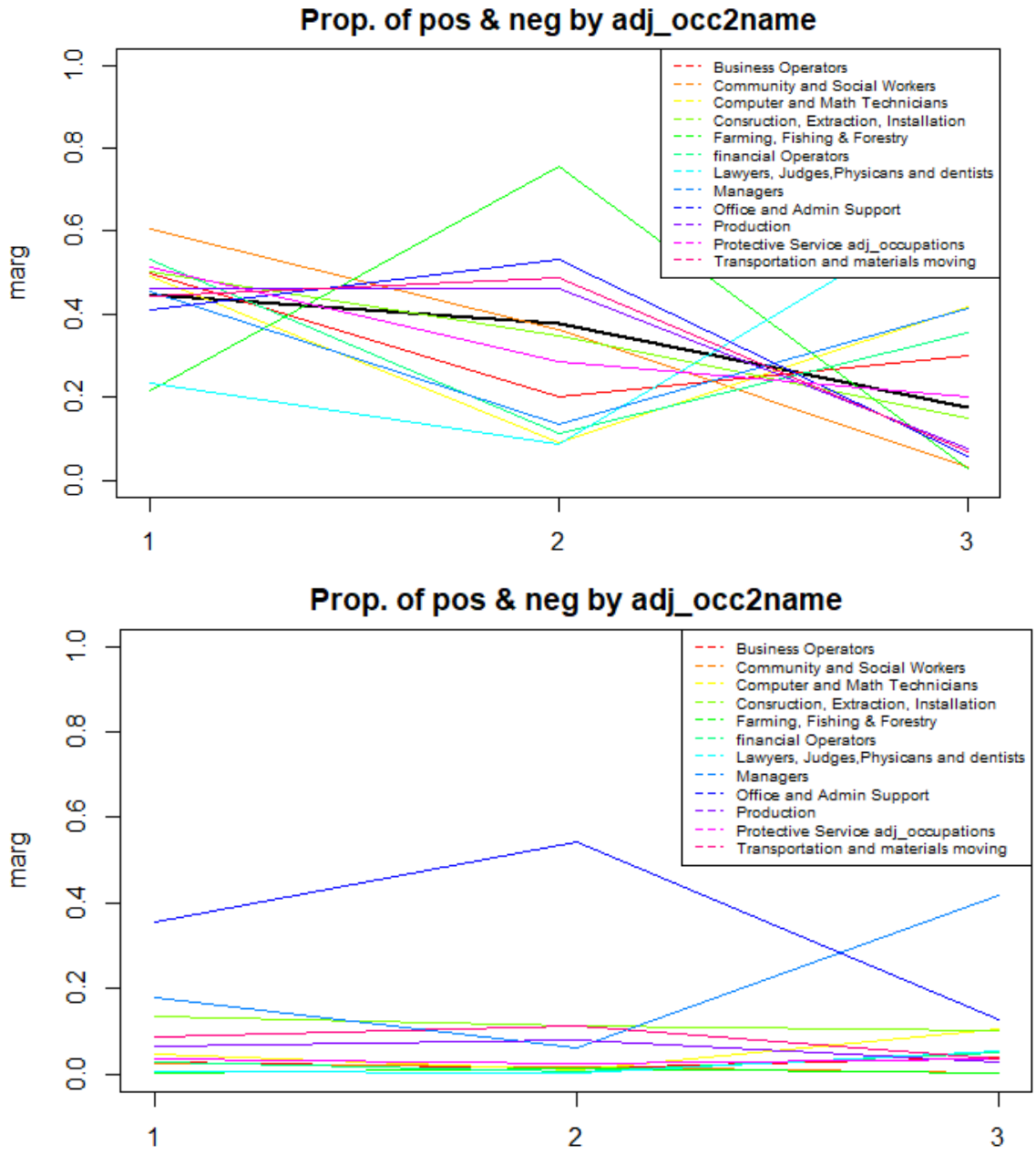


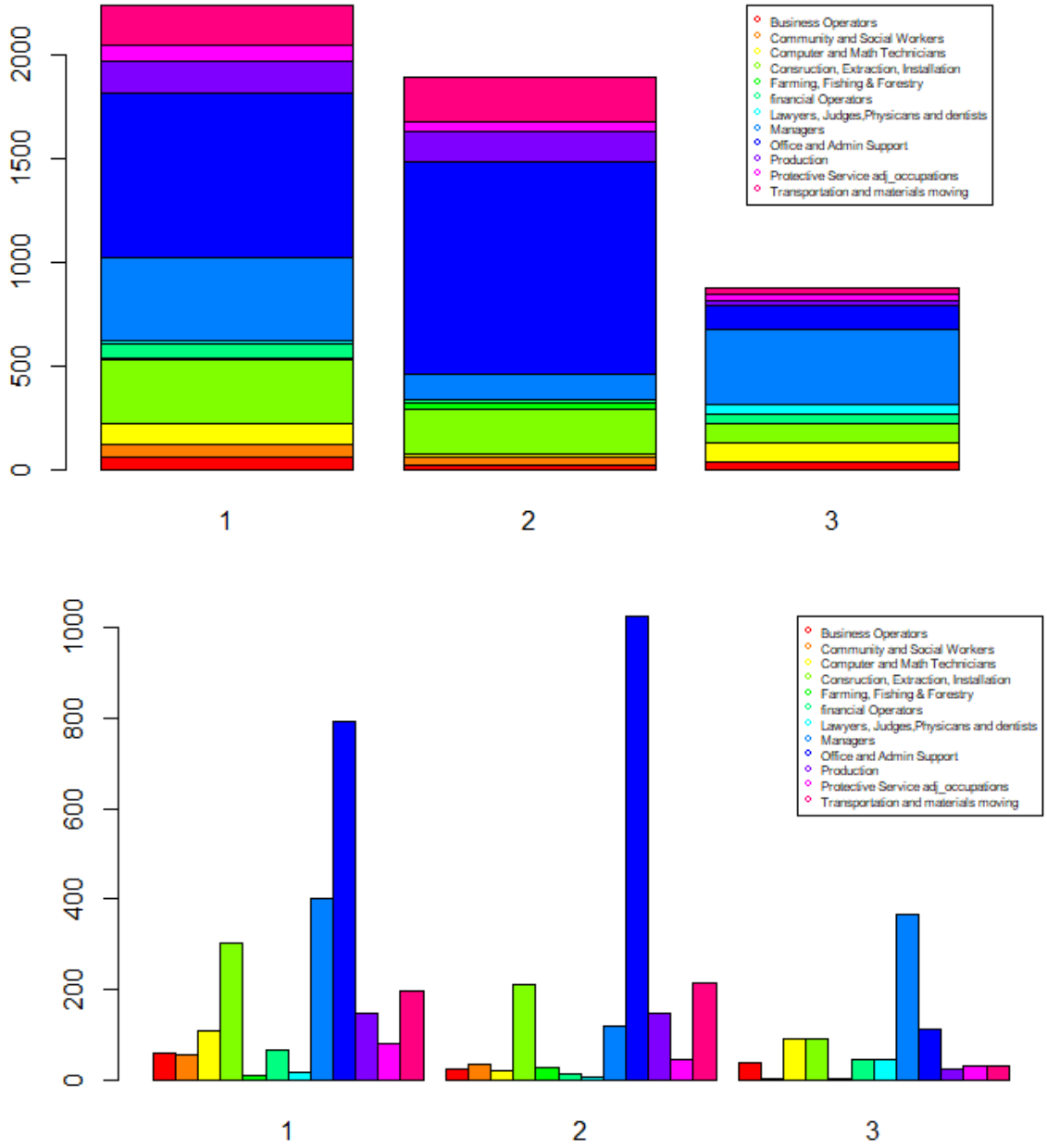
The three clusters have very similar values for this variable, cluster two has a slightly higher percentage of individuals whose family type is related subfamily, secondary individual or unrelated subfamily, but it's not a big difference.

6.2.13 Ajd_ind

This variable has too many values and the plots obtained are quite impossible to read. We had a similar problem with bpl, but in that case grouping some of them together was simpler and made more sense than in this case. For those reasons we have decided to not include the plots for this variable. However, they can still be obtained with our script, and we have some similar information from Adj_occ2name.

6.2.14 Adj_occ2name





Cluster one and two's most frequent value is Office and Admin Support with cluster two having the highest number of them, whereas cluster three's most frequent value is actually Managers, with more than 40% of their individuals being managers..

Cluster three also has the highest percentage of Lawyers, Judges, Physicians and dentists, business operators, and financial operators, and the lowest amount of community and social, production and transportation workers.

These results are probably related to the ones we saw in variables Sch and Educ99.

6.3 Clusters characterization

We have seen and commented on the values each variable takes for each cluster, now it's time to draw some conclusions about them. It might be a bit confusing but we have decided to explain cluster two and cluster three before explaining cluster one.

6.3.1 Cluster 2

Cluster two is the one with the lowest mean Incwage and Hrwage, with its mean Incwage being lower than half of the global mean.

Cluster two is the one with the highest percentage of females, it's the only one with more females than males. In terms of race, it is also the one with the highest amount of Hispanic individuals and the lowest percentage of white non hispanic individuals, this is also reflected in the Bpl variable, cluster two is the one with the highest amount of individuals with Mexico as their birthplace, and it's the one with the lowest amount of people with U.S as their birthplace, so it's the cluster with the highest amount of people with races that are not white.

Cluster two is also the one with the lowest level of education, with most of their individuals deciding to finish their education right after or before finishing high school.

If we talk about their occupations, cluster two is the one with the highest percentage of individuals working in office and admin support, with most of the individuals working on this occupation.

About their marital status, cluster two is actually the one with the highest percentage of individuals who are never divorced, never married, separated, widowed and married with an absent spouse.

There are some other variables where cluster two is quite different from the rest, but we have considered these differences to not be as relevant because of the number of individuals affected by it. To give an example of this, for the variable relate, cluster two is the one with the highest percentage of individuals with children as their value. But the amount of individuals with this value is really small, so we considered it to not be so important.

6.3.2 Cluster 3

Cluster three is the one with the highest mean Incwage and Hrwage, with their mean being more than double than the global mean. Cluster three is the only one with outliers for the Incwage variable, with a considerable amount of them, this could be influencing the mean. Still, the rest of the values are still higher than the ones from other clusters.

Cluster three has more than double males than females, with almost 70% of its individuals being males.

In terms of race cluster three has the lowest percentage of hispanic individuals, and the highest percentage of white non hispanic individuals. This is also reflected in the Bpl variable, cluster three is the one with the lowest percentage of individuals with Mexico as their birthplace.

About their marital status cluster three is the one with the lowest percentage of individuals who never married and married individuals with an absent spouse.

Cluster three is the one with the highest level of education, it has the highest percentage of individuals with a doctoral, masters, advanced or bachelor's degree, it also has the lowest percentage of individuals that decided to finish their education right after or before finishing highschool.

In terms of their occupations, cluster three's most frequent occupation is Managers, it has the highest percentage of them among all the clusters, with 40% of their individuals having this occupation. It also has almost no community and social workers, and compared to the rest of clusters, very little office and admin support workers.

6.3.3 Cluster 1

Cluster one is the one with the highest number of individuals. It's also the most balanced one in terms of the sex of the individuals.

For most of the variables cluster one is actually kind of an inbetween between cluster two and cluster three.

	Incwage	Sex	Race	Educ99	Occupation
Low Income Cluster (Cluster 2)	Lowest mean income	Highest % of female individuals	Highest % of non-white people	Lowest education level	High % of Office & Admin support
High Income Cluster (Cluster 3)	Highest mean income	Highest % of male individuals	Lowest % of Hispanic	Highest education level	High % of Managers
Middle Income Cluster (Cluster 1)	Average income almost equal to global inc wage	Balanced % of male and female	Race is balanced	In between cluster 1 and 2	High % of Office & Admin support

7. Conclusions

Based on our analysis using PCA and clustering techniques, we have drawn several conclusions about the existence of a gender pay gap and the variables that affect it. Our clustering results show that Cluster 2, which has the highest percentage of females and lowest mean Incwage and Hrwage, represents the group with the most significant pay gap. This cluster also has the highest percentage of Hispanic individuals, lowest level of education, and highest percentage of individuals working in office and admin support. Additionally, Cluster 3 has the highest mean Incwage and Hrwage, highest percentage of males, highest level of education and highest percentage of white non hispanic individuals.

Our PCA analysis also confirms the existence of a gender pay gap, as we have observed that males are more likely to earn more money compared to females. Moreover, owning characteristics such as being male, white, and married seems to be more considered by society, which leads to a better salary. We have also found that individuals with higher degrees have better-paying jobs, and there is a clear correlation between salary and working hours, but no relation between age and salary.

Overall, our results suggest that certain variables such as gender, race, education level, and occupation play a significant role in determining the gender pay gap. We can conclude that the gender pay gap does exist in our dataset. Our findings suggest that individuals who are male, white, and married are more likely to earn higher salaries compared to those who do not possess these characteristics, indicating that society policies and norms may have a bias towards these groups. Additionally, we observed that higher education levels are associated with better paying jobs, indicating that education may play a role in the gender pay gap.

In conclusion, our analysis suggests that the pay gap exists in our dataset, and that gender and education may be important variables in explaining the gap, but further analysis is needed to fully understand the factors contributing to the pay gap.

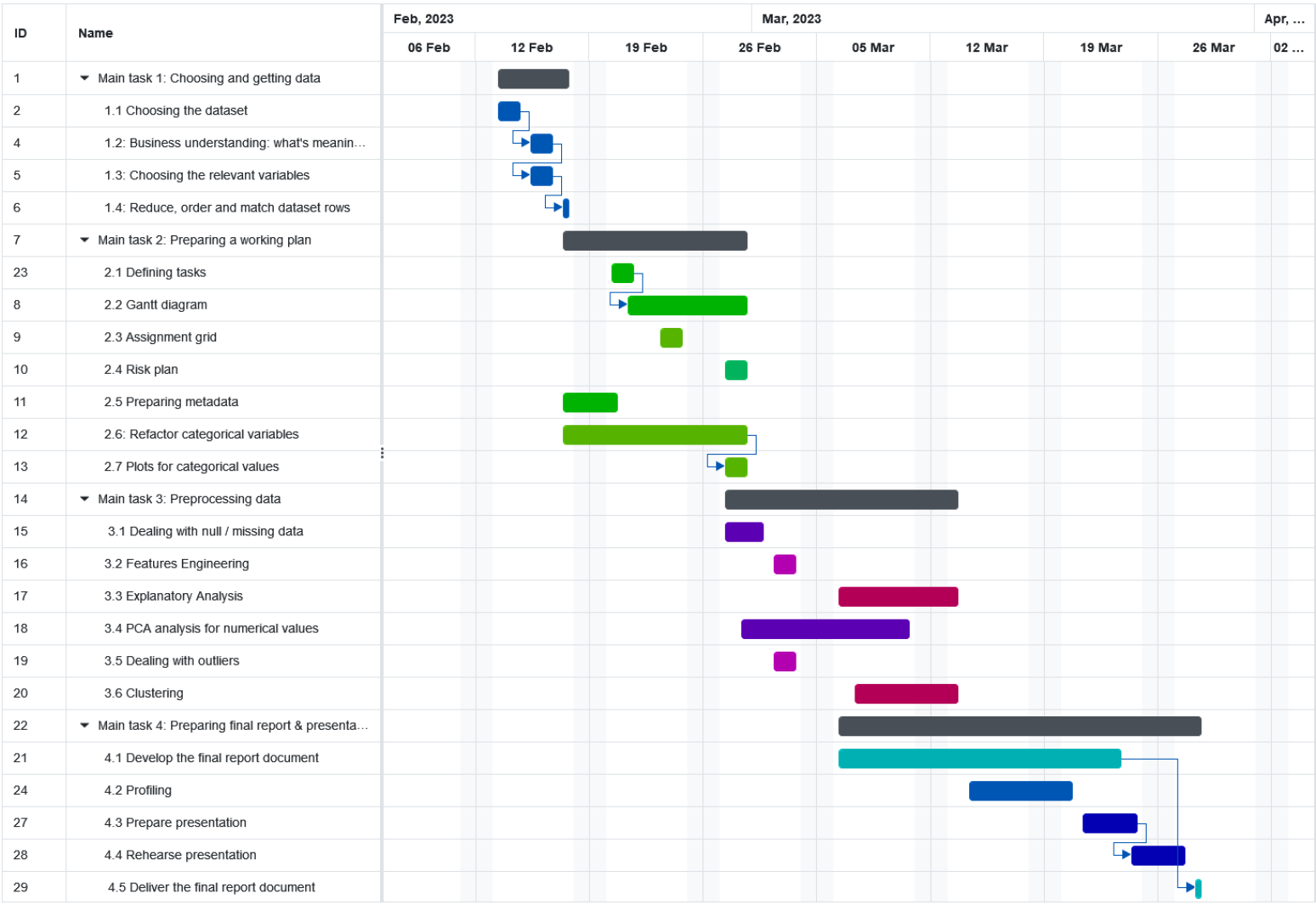
8. Working plan

Throughout the duration of the project, we can note that the execution of the project has remained consistent with the working plan and the Gantt Diagram that was presented at the outset of the project. While it is true that there were a few deviations from the initial plan, it is worth emphasizing that these deviations were minor and infrequent, and any necessary adjustments that were made were executed with the utmost punctuality.

8.1. Gantt diagram

Initial:

Final:



8.2. Working plan

	Miguel	Alessandro	Laia	Fujie	Joan
Main task 1: Choosing and getting data					
Task 1.1: Choosing the dataset	X	X	X	**X**	X
Task 1.2: Business understanding	X	X	X	X	**X**

Task 1.3: Choosing the relevant variables	X	**X**	X	X	X
Task 1.4: Reduce, order and match dataset rows	**X**				
Main task 2: Preparing a working plan					
Task 2.1: Gantt diagram	X				**X**
Task 2.2: Assignment grid	**X**	X	X	X	X
Task 2.3: Risk plan	**X**		X	X	
Task 2.4: Preparing metadata		**X**			X
Task 2.5: Refactor categorical variables	**X**	X	X	X	X
Task 2.6: Plots for categorical values	**X**	X	X	X	X
Main task 3: Preprocessing data					
Task 3.1: Dealing with null / missing data	X			**X**	
Task 3.2: Features Engineering	**X**	X	X	X	X
Task 3.3: Explanatory Analysis	X	**X**	X	X	X
Task 3.4: PCA analysis for numerical values	X	X	X	**X**	X

Task 3.5: Dealing with outliers	X	X	**X**	X	X
Task 3.6: Clustering	X	X	X	**X**	X
Main task 4: Preparing final report & presentation					
Task 4.1: Develop the final report document	X	X	**X**	X	X
Task 4.2: Profiling	**X**	X	X	X	X
Task 4.3: Predictive modeling/logistic regression		**X**			X
Task 4.4: Prepare presentation	X	X	X	X	**X**
Task 4.5: Rehearse presentation	X	**X**	X	X	X
Task 4.6: Deliver the final report document	X	X	**X**	X	X

8.3. Risk plan

During the course of the project, there were a few instances where it became necessary to apply certain points that had been identified and laid out in the risk plan that was developed prior to the initiation of the project. Specifically, these points can be summarized as follows:

- Huge amount of null values: we encountered a significant issue with an exceedingly large number of null values within a particular column of the dataset. As per the risk plan, we proceeded with the pre-established strategy to tackle this issue, which involved deleting the entire column from the dataset.
- Datasets has conflicts: we encountered conflicts within certain columns of the datasets, which necessitated replacement in order to ensure the integrity of the data. Again, in accordance with the risk plan, we proceeded with the pre-planned strategy that had been developed in the initial grid.
- Late delivery: we were very careful with the deliveries dates and we scheduled everything with extra time.

- A member is not working: there were instances in which a team member was unable to contribute to the project due to unforeseen circumstances. As per the risk plan, we applied pressure to ensure that the individual in question was kept in touch with the project and remained on track with their designated responsibilities.
- Not up to date: related with the previous point, redefinition of tasks was necessary.

Overall, while it is true that there were a few minor conflicts that arose during the project, it is worth emphasizing that these issues were managed and in accordance with the pre-established risk plan. As a result, no major problems occurred during the course of the project.

Risk	**How to prevent**	**How to manage**
A team member leaves the course	All tasks have at least two members assigned	Pending work reassigned to re-balance efforts
Late deliveries	Plan with extra time in case of any issues	Work together to finish it in time and talk it out with the professor in case it's not possible
Dataset has conflicts	Look for another dataset in which the values are more descriptive and well introduced	Change some variables or replace them from the original dataset
Huge amount of nulls values	Check the dataset and look up for the percentage of each variable	Depending on the percentage, we will delete or set the value as not available
PCA plots not valid	Make sure that we have proceeded correctly and the results are valid	Present the plots to the professor
Disputes with the teams members	Distribute tasks in an equal way and argue effectively, with humility	Maintain calm and promote communication with members to solve it
Not useful variables	Make an inspection to each variable and look up for any misunderstanding relation	Depending on the situation, we should delete, change or transform it to solve the problem
A member is not working	Make sure the task planning is known and consider the schedule	Make group pressure, so the member can reincorporate and also give some help in case that is needed
Tasks not well-defined or distribute	Consider the schedule and organize it with all the members to see how much time each task will take and allocate them	See how much time each person has to work and configure it so solve the wrong distributed task.

Not up to date	Work every week and communicate with teammates	Take account of each member schedule and redefine it in case that is not completing the tasks