

**Report**

***United States census income data***

Data Mining 2022-2023 Q2

**Authors:**

Pol Casacuberta Gil

He Chen

Eric Hurtado

Tommaso Patriti

Alexandru-Ilie Popa

**Facultat d’Informàtica de Barcelona**

[**1.Initial working plan 3**](#_ayvm2zlxhdie)

[**2.Formal description of data structure and 6**](#_uok16r2tim1x)

[**metadata 6**](#_f0jucwvtqdpc)

[2.1 Numerical Variable Metadata 6](#_779f8tdo112d)

[2.2 Categorical Variable Metadata 7](#_fxaspz6nx59e)

[**3.Detailed description of preprocessing and data preparation 22**](#_crd8kfuxy4je)

[3.1 Feature selection 25](#_17e0pu5r6o7l)

[3.2 Statistical sampling 25](#_415by2m2lmq7)

[3.3 Missing data 26](#_tj88n8oml1r9)

[**4.Basic statistical descriptive analysis 29**](#_skf1t0u2ah8p)

[4.1 Univariate analysis 31](#_fkg9b23kilgj)

[4.2 Bivariate plots 82](#_j0o7639bzg30)

[4.3 Conclusion 88](#_ypqqc2tksfz6)

[**5.PCA analysis for numerical variables 89**](#_kprj4huxbpfa)

[5.1 Scree plot 89](#_jy7bfp2l8jqi)

[5.2 Factorial map 90](#_vzll0l8u3oif)

[5.2.1 Individuals plot 91](#_vkwe72nafj1w)

[5.2.2 Common projection of numerical variables and modalities 93](#_64u7unaajmp5)

[5.2.3 Interpretation of relationships among the observed variables 95](#_kdwp4znchzju)

[5.2.4 Conclusion 95](#_scwtwdvzrvet)

[**6.Clustering 96**](#_xamnxbxbpj2t)

[6.1 Clustering Method & Resulting Dendrogram 96](#_utqv5nipm5ox)

[6.1.1 Dendrogram method="ward.D" 96](#_jyvxzca0h3bg)

[6.1.2 Dendrogram method="ward.D2", dissimilarity matrix 97](#_b40nt7my07ka)

[6.2 Clusters Size 98](#_qwn5bl6n65sv)

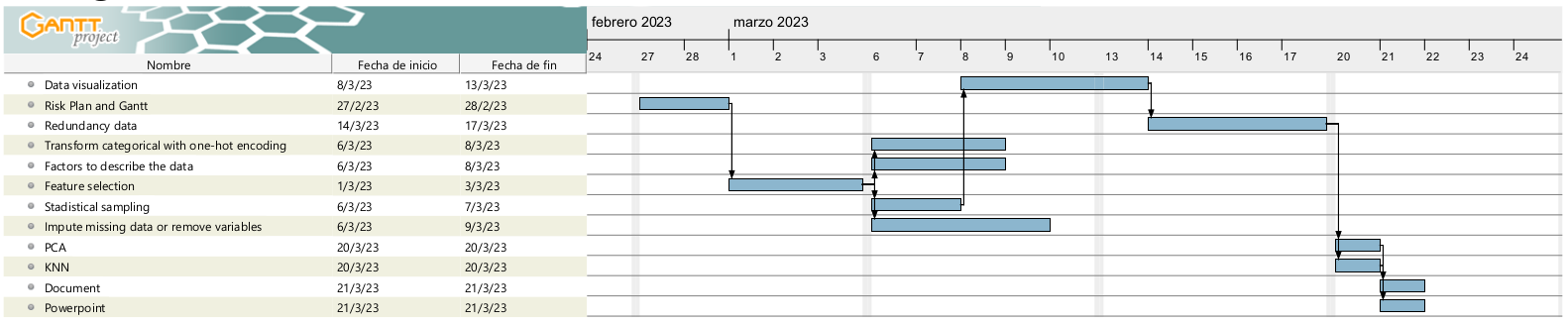
[6.3 Cluster analysis and Conclusion 98](#_2azu6oykbkus)

# 

# 1.Initial working plan

| Participant | Alex | Pol | Tommaso | Eric | He |
| --- | --- | --- | --- | --- | --- |
| 1. Risk Plan and Gantt(2) D2 |  | X |  |  | X |
| 2. Preprocessing | X |  |  |  |  |
| 2.1 Feature selection (2) (D1) |  |  |  |  | X |
| 2.2 Stadistical sampling (D2) | X |  |  |  |  |
| 2.3 Redundacy data (5) (D4) |  |  | X |  |  |
| 2.4 Technical review of data (D1) |  |  |  |  |  |
| 2.5 Transform categorical into one-hot encoding (4) (D3) |  | X |  |  |  |
| 2.6 Impute missing data (2) (D4) |  |  |  |  | X |
| 2.7 Data visualization (1) (D5) | X |  |  |  |  |
| 2.8 Factors to describe (3) (D3) | X |  |  | X |  |
| 3. PCA (4) | X | X |  |  |  |
| 4. KNN (5) |  |  | X |  |  |
| 5. Clustering (5) |  |  | X |  |  |
| 6. Profiling (3) |  |  |  | X |  |
| 7. Document (word style) | X | X | X | X | X |
| 8. Powerpoint | X | X | X | X | X |
|  |  |  |  |  |  |
| DURATION: from shortest to longest D1-D5 |  |  |  |  |  |

**Workflow grid:**



**Risk: Data security breach**

**Prevention**

Ensure all team members have proper security clearance and access levels to sensitive data. Use encryption technology and secure file sharing platforms through Github with SSH protocols.

**Management**

Immediately notify the group. Identify the cause of the breach and take steps to prevent it from happening again. Make sure to make the repository private and not share confidential documents with other groups.

**Risk: Unavailability of key team members**

**Prevention**

Have a clear project timeline and assign clear roles and responsibilities to each team member. Identify backup team members who can fill in if someone is unavailable. Hold regular team meetings and communicate frequently to ensure everyone is aware of their responsibilities and progress.

**Management**

If a team member becomes unavailable, immediately identify a backup team member to take their place. Reallocate responsibilities as necessary to ensure the project stays on track. Communicate any changes to the group.

**Risk: Inaccurate or incomplete data**

**Prevention**

Establish a clear data quality control process, including data cleaning and validation. Have a clear understanding of the data source and any limitations or biases.

**Management**

Address any errors or omissions in the data as soon as they are discovered. Re-evaluate the data quality control process and make any necessary adjustments.

**Risk: Scope creep**

**Prevention**

Clearly define the project scope and deliverables. Have a clear understanding of the project objectives and priorities. Regularly review progress against the project plan.

**Management**

Any changes to the project scope must be reviewed and approved by the group. If a change is approved, update the project plan and communicate the changes to the team. Manage resources and timelines to ensure any scope changes do not impact the overall project delivery.

**Risk: Technical failure**

**Prevention**

Use reliable hardware and software. Have a backup plan for data storage and disaster recovery. Regularly test the technology and systems being used.

**Management**

Immediately identify the cause of the technical failure and take steps to address it. Implement the backup plan for data storage and disaster recovery and store data on Github.

# 

# 2.Formal description of data structure and

# metadata

## 2.1 Numerical Variable Metadata

| **Name** | **Type** | **Range** | **Units** | **Missing percent** | **Meaning** |
| --- | --- | --- | --- | --- | --- |
| age | numerical | [0, 90] | years | 0 | respondent’s age |
| wage.per.hour | numerical | [0, 8000] | dollars | 0 | respondent’s wage per hour |
| capital.gains | numerical | [0, 99999] | dollars | 0 | respondent’s capital gains |
| capital.losses | numerical | [0, 4356] | dollars | 0 | respondent’s capital losses |
| dividends.from.stocks | numerical | [0, 99999] | dollars | 0 | respondent’s earnings by dividends from stock |
| num.persons.worked.for.employer | numerical | [0, 6] |  | 0 | respondent’s employees |
| weeks.worked.in.year | numerical | [0, 52] | weeks | 0 | respondent’s weeks worked in a year |
| instance.weight | numerical | [40.67, 11352.5] |  | 0 | respondent’s capacity units that each instance type would contribute |

## 2.2 Categorical Variable Metadata

| **Variable** | **Type** | **Modalities** | **Missing percent** | **Meaning** |
| --- | --- | --- | --- | --- |
| class.of.worker | categorical | Not considered  Self-employed-not incorporated  Local government  Private  Self-employed-incorporated  State government  Never worked  Federal government  Without pay | 0 | respondent’s class of worker |
| detailed.industry.recode | categorical | Not considered  Other Agriculture  Educational Services  Other Professional Services  Banking and Other Finance  Retail Trade  Fabricated metal  Construction  Wolesale Trade  Other Public Administration  Business Services  Machinery, except electrical  Hospitals  Primary metals  Professional and photographic equipment, and watches  Transportation  Aircraft and parts  Personal Services, Except Private Household  Paper and allied products  Private Household Services  Communications  Printing, publishing and allied industries  Rubber and miscellaneous plastics products  Motor vehicles and equipment  Electrical machinery, equipment, and supplies  Apparel and other finished textile products  Food and kindred products  Mining  Armed Forces last job, currently unemployed  Textile mill products  Petroleum and coal products  Stone clay, glass, and concrete product  Insurance and Real Estate  National Security and Internal Affairs  Health Services, Except Hospitals  Agriculture Service  Entertainment and Recreation Services  Repair Services  Social Services  Administration of Human Resource Programs  Utilities and Sanitay Services  Furniture and fixtures  Chemicals and allied products  Other transportation equipment  Leather and leather products  Lumber and wood products, except furniture  Toys, amusements, and sporting goods  Miscellaneous and not specified manufacturing industries  Justice, Public Order and Safety  Forestry and Fisheries  Tobacco manufactures | 0 | respondent’s industry |
| detailed.occupation.recode | categorical | Not considered  Farm Operators and Managers  Teachers, Except College and University  Technicians, Except Health, Engineering, and Science  Other Executive, Administrators, and Managers  Financial Records, Processing Occupations  Food Service Occupations  Teachers, College and University  Construction Trades  Sales Workers, Retail and Personal Services  Other Administrative Support Occupations, Including Clerical  Secretaries, Stenographers, and Typists  Cleaning and Building Service Occupations  Management Related Occupations  Other Precision Production Occupations  Machine Operators and Tenders, Except Precision  Freight, Stock and Material Handlers  Engineering and Science Technicians  Mechanics and Repairers  Lawyers and Judges  Private Household Service Occupations  Public Administration  Protective Service Occupations  Motor Vehicle Operators  Farm Workers and Related Occupations  Health Assessment and Treating Occuaptions  Other Transportation Occupations and Material Moving  Personal Service Occupations  Armed Forces last job, currently unemployed  Other Handlers, Equipment Cleaners, and Laborers  Sales Representatives, Finance, and Business Service  Construction Laborer  Supervisors - Administrative Support  Supervisors and Proprietors, Sales Occupations  Computer Equipment Operators  Health Technologists and Technicians  Other Professional Specialty Occupations  Mathematical and Computer Scientists  Sales Representatives, Commodities, Except Retail  Health Service Occupations  Fabricators, Assemblers, Inspectors, and Samplers  Health Diagnosis Occupations  Engineers  Mail and Message Distributing  Natural Scientists  Forestry and Fishing Occupations  Sales Related Occupations | 0 | respondent’s occupation |
| education | categorical | Children  High school graduate  Bachelors degree(BA AB BS)  Some college but no degree  Associates degree-occup /vocational  11th grade  5th or 6th grade  Masters degree(MA MS MEng MEd MSW MBA)  10th grade  7th and 8th grade  9th grade  12th grade no diploma  Prof school degree (MD DDS DVM LLB JD)  Doctorate degree(PhD EdD)  1st 2nd 3rd or 4th grade  Associates degree-academic program  Less than 1st grade | 0 | respondent’s education |
| marital.stat | categorical | Never married  Married-civilian spouse present  Divorced  Widowed  Separated  Married-spouse absent  Married-A F spouse present | 0 | respondent’s civil status |
| major.industry.code | categorical | Not considered  Agriculture  Education  Other professional services  Finance insurance and real estate  Retail trade  Manufacturing-durable goods  Construction  Wholesale trade  Public administration  Business and repair services  Hospital services  Transportation  Personal services except private HH  Manufacturing-nondurable goods  Private household services  Communications  Mining  Armed Forces  Medical except hospital  Entertainment  Social services  Utilities and sanitary services  Forestry and fisheries | 0 | respondent’s major industry |
| major.occupation.code | categorical | Not considered  Farming forestry and fishing  Professional specialty  Technicians and related support  Executive admin and managerial  Adm support including clerical  Other service  Precision production craft & repair  Sales  Machine operators assmblrs & inspctrs  Handlers equip cleaners etc  Private household services  Protective services  Transportation and material moving  Armed Forces | 0 | respondent’s major occupation |
| race | categorical | Black  White  Amer Indian Aleut or Eskimo  Asian or Pacific Islander  Other | 0 | respondent’s race |
| hispanic.origin | categorical | All other  Mexican-American  Mexican (Mexicano)  Central or South American  UnknownHispanicOrigin  Other Spanish  Puerto Rican  Cuban  Do not know  Chicano | 0 | respondent’s origin |
| sex | binary | Female  Male | 0 | respondent’s sex |
| full.or.part.time.employment.stat | categorical | Children or Armed Forces  Full-time schedules  PT for non-econ reasons usually FT  Not in labor force  Unemployed full-time  Unemployed part- time  PT for econ reasons usually PT  PT for econ reasons usually FT | 0 | respondent’s stats of employment |
| tax.filer.stat | categorical | Nonfiler  Joint both under 65  Single  Head of household  Joint both 65+  Joint one under 65 & one 65+ | 0 | respondent’s tax filer stat |
| detailed.household.and.family.stat | categorical | Child <18 never marr not in subfamily  Householder  Spouse of householder  Child 18+ never marr Not in a subfamily  Child under 18 of RP of unrel subfamily  Other Rel 18+ never marr not in subfamily  Nonfamily householder  Child 18+ ever marr RP of subfamily  Other Rel 18+ ever marr not in subfamily  Secondary individual  Grandchild <18 never marr child of subfamily RP  RP of unrelated subfamily  Grandchild 18+ never marr not in subfamily  Other Rel 18+ spouse of subfamily RP  In group quarters  Other Rel 18+ ever marr RP of subfamily  Child 18+ ever marr Not in a subfamily  Other Rel <18 never marr not in subfamily  Child 18+ spouse of subfamily RP  Spouse of RP of unrelated subfamily  Grandchild <18 never marr not in subfamily  Child 18+ never marr RP of subfamily  Other Rel <18 never marr child of subfamily RP  Child <18 never marr RP of subfamily  Other Rel 18+ never marr RP of subfamily  Other Rel <18 ever marr RP of subfamily  Grandchild 18+ ever marr not in subfamily  Child <18 ever marr not in subfamily  Grandchild 18+ ever marr RP of subfamily  Child <18 ever marr RP of subfamily  Grandchild 18+ spouse of subfamily RP  Other Rel <18 never married RP of subfamily | 0 | respondent’s household and family stat |
| detailed.household.summary.in.household | categorical | Child under 18 never married  Householder  Spouse of householder  Child 18 or older  Nonrelative of householder  Other relative of householder  Group Quarters- Secondary individual  Child under 18 ever married | 0 | respondent’s household summary |
| country.of.birth.father | categorical | United-States  Dominican-Republic  Mexico  Taiwan  Canada  UnknownFatherCountry  China  Peru  Ireland  Haiti  Cuba  Italy  Portugal  Poland  Nicaragua  El-Salvador  England  Puerto-Rico  India  Philippines  France  Iran  Cambodia  Outlying-U S (Guam USVI etc)  Honduras  Scotland  Greece  Germany  Guatemala  Ecuador  Japan  Laos  Thailand  South Korea  Yugoslavia  Hungary  Vietnam  Jamaica  Columbia  Holand-Netherlands  Trinadad&Tobago  Hong Kong  Panama | 0 | respondent’s country birth of father |
| country.of.birth.mother | categorical | United-States  Canada  Mexico  Taiwan  China  UnknownMotherCountry  Ireland  England  Haiti  Cuba  Italy  Dominican-Republic  Outlying-U S (Guam USVI etc)  Poland  Germany  Nicaragua  Japan  El-Salvador  Peru  India  Iran  Puerto-Rico  Honduras  Philippines  South Korea  Greece  Guatemala  Ecuador  Laos  Thailand  Scotland  Hungary  Vietnam  Jamaica  Columbia  France  Portugal  Yugoslavia  Cambodia  Hong Kong  Panama  Holand-Netherlands  Trinadad&Tobago | 0 | respondent’s country birth of mother |
| country.of.birth.self | categorical | United-States  Mexico  Taiwan  China  Ireland  Canada  Haiti  Dominican-Republic  UnknownSelfCountry  Outlying-U S (Guam USVI etc)  Poland  Germany  Nicaragua  England  Peru  India  Iran  Puerto-Rico  Cuba  Philippines  Greece  Guatemala  Ecuador  Japan  El-Salvador  Laos  Thailand  South Korea  Vietnam  Jamaica  Columbia  Italy  Honduras  France  Yugoslavia  Scotland  Hong Kong  Panama  Trinadad&Tobago  Portugal  Cambodia  Holand-Netherlands  Hungary | 0 | respondent’s country birth |
| citizenship | categorical | Native- Born in the United States  Native- Born abroad of American Parent(s)  Foreign born- Not a citizen of U S  Foreign born- U S citizen by naturalization  Native- Born in Puerto Rico or U S Outlying | 0 | respondent’s citizenship |
| veterans.benefits | categorical | UnknownVeteranBenefits  2  1 | 0 | respondent’s veteran benefits |
| income | categorical | Less than 50000  Greater than 50000 | 0 | respondent’s income |

# 3.Detailed description of preprocessing and data preparation

In the following table, there is a summary of the features information:

| **#** | **Feature Name** | **Description** | **Type** | **% Missing data** |
| --- | --- | --- | --- | --- |
| 1 | age | The age of the individual in years. | Numeric | 0 |
| 2 | class of worker | The type of work arrangement or employer for which the individual works. | Categorical | 50.242328 |
| 3 | detailed industry recode | A detailed code indicating the specific industry in which the individual is employed. | Categorical | 0 |
| 4 | detailed occupation recode | A detailed code indicating the specific occupation in which the individual is employed. | Categorical | 0 |
| 5 | education | The highest level of education completed by the individual. | Categorical | 0 |
| 6 | wage per hour | The hourly wage rate for the individual's job. | Numeric | 0 |
| 7 | enroll in edu inst last wk | Whether or not the individual was enrolled in an educational institution during the previous week. | Categorical | 93.6949625 |
| 8 | marital stat | The marital status of the individual. | Categorical | 0 |
| 9 | major industry code | A broad code indicating the major industry in which the individual is employed. | Categorical | 50.4623527 |
| 10 | major occupation code | A broad code indicating the major occupation in which the individual is employed. | Categorical | 50.4623527 |
| 11 | race | The individual's race. | Categorical | 0 |
| 12 | hispanic origin | Whether or not the individual identifies as Hispanic or Latino. | Categorical | 0.4380447 |
| 13 | sex | The individual's gender. | Categorical | 0 |
| 14 | member of a labor union | Whether or not the individual is a member of a labor union. | Categorical | 90.4452118 |
| 15 | reason for unemployment | The reason why the individual is currently unemployed. | Categorical | 96.9577442 |
| 16 | full or part-time employment stat | Whether the individual is employed full-time or part-time. | Categorical | 0 |
| 17 | capital gains | The amount of capital gains earned by the individual during the year. | Numeric | 0 |
| 18 | capital losses | The amount of capital losses incurred by the individual during the year. | Numeric | 0 |
| 19 | dividends from stocks | Amount of dividends earned from stocks or mutual funds during the year for each individual | Numeric | 0 |
| 20 | tax filer stat | Whether or not the individual is required to file a tax return. | Categorical | 0 |
| 21 | region of previous residence | The region of the United States where the individual lived one year ago. | Categorical | 92.0946457 |
| 22 | state of previous residence | The state where the individual lived one year ago. | Categorical | 92.449492 |
| 23 | detailed household and family stat | A detailed code indicating the household and family status of the individual. | Categorical | 0 |
| 24 | detailed household summary in household | A detailed code indicating the type of household in which the individual resides. | Categorical | 0 |
| 25 | instance weight | A weight assigned to each individual in the dataset to adjust for sampling and non-response biases. | Continuous | 0 |
| 26 | migration code-change in msa | Whether the individual moved to a different metropolitan statistical area (MSA) between the previous year and the current year. | Categorical | 50.7269839 |
| 27 | migration code-change in reg | Whether the individual moved to a different region of the United States between the previous year and the current year. | Categorical | 50.7269839 |
| 28 | migration code-move within reg | Whether the individual moved within the same region of the United States between the previous year and the current year. | Categorical | 50.7269839 |
| 29 | live in this house 1 year ago | Whether the individual lived in the same house one year ago. | Categorical | 50.7269839 |
| 30 | migration prev res in sunbelt | Whether the individual moved from a state in the Sunbelt region of the United States between the previous year and the current year. | Categorical | 92.0946457 |
| 31 | num persons worked for employer | The number of people who worked for the individual's employer during the year. | Continuous | 0 |
| 32 | family members under 18 | The number of family members under the age of 18 living in the same household as the individual. | Continuous | 72.2884079 |
| 33 | country of birth father | The country of birth of the individual's father. | categorical | 3.3645244 |
| 34 | country of birth mother | The country of birth of the individual's mother. | categorical | 3.0668144 |
| 35 | country of birth self | The country of birth of the individual. | categorical | 1.7005558 |
| 36 | citizenship | Whether the individual is a U.S. citizen, a non-citizen with a green card, or a non-citizen without a green card. | categorical | 0 |
| 37 | own business or self-employed | Whether the individual owns a business or is self-employed. | categorical | 0 |
| 38 | fill inc questionnaire for veteran's admin | Whether the individual filled out an income questionnaire for the Veterans Administration. | categorical | 99.0056284 |
| 39 | veterans benefits | Whether the individual receives veterans benefits. | categorical | 0 |
| 40 | weeks worked in year | The number of weeks the individual worked during the year. | numerical | 0 |
| 41 | year | The year in which the data was collected. | categorical | 0 |

## 3.1 Feature selection

First of all, we have analyzed the meaning of the variables. The intention is to find columns that do not contribute any information to the topic we are analyzing. The result of this first analysis is that we have removed the variable "year", which represents The year in which the data was collected, which has nothing to do with the topic we are analyzing.

## 3.2 Statistical sampling

In our database, we have 199,523 cases. As it is a fairly large number, we decided to only take 10%, which is 20,000 cases. To do this, we applied random sampling.

## 3.3 Missing data

Our database has used different ways to represent missing data, including "Not in universe", "Not in universe or children", "?", "Not in universe under 1 year old". After analyzing them, we realized that these values either mean that we don't really know the information or the question is not relevant to the individual's situation. For example, it doesn't make sense to ask if a child is working or not. In any case, these are values that do not provide any information. For convenience, we have converted all these values to "NA".

Next, we calculated the percentage of missing data for each feature (which you can see in the table above). In summary, we can classify variables into three groups based on their percentage of missing values:

1. % Missing values <= 10%: hispanic origin, country of birth father, country of birth mother, country of birth self
2. 90% > % Missing values >= 50%: class of worker, major industry code, major occupation code, migration code-change in msa, migration code-change in reg, migration code-move within reg, live in this house 1 year ago, family members under 18
3. % Missing values >= 90%: enrolled in edu inst last wk, member of a labor union,

We start the analysis with group 1: Since the % of missing data is very low in this group, we will try to impute and give value to the cells with NA, taking into account that this value cannot affect subsequent analyses. An interesting characteristic is that all the features in this group are related to the individual's or close persons' country of origin.

| **#** | **Feature** | **Analysis** | **Results** |
| --- | --- | --- | --- |
| 12 | hispanic origin | The % of missing data (0.4) is very low in this case. However, we have encountered a difficulty in finding an imputation method that fits in this case. To know its exact value, we have to pursue the ancestors of the individual. | Created: "UnknownHispanicOrigin" category. |
| 33/34/35 | country of birth father / country of birth mother / country of birth self | These three features were analyzed together because they are closely related. If the father was born in country X, there is a high probability that the mother was also born there, and the same goes for the child. At first, we considered imputing them using the hot desk method, but this strategy can be unreliable when we know nothing about any of these three features. In the end, we decided to create a new category. | New categories created:  "UnknownFatherCountry",  "UnknownMotherCountry",  "UnknownSelfCountry". |

We consider this feature to be very important since it allows us to know which type of work is better paid. However, due to its high percentage of missing data, it is impossible to impute it without having an impact on the results. Therefore, we decided to create a new category.

| **#** | **Feature** | **Analysis** | **Results** |
| --- | --- | --- | --- |
| 2 | class of worker | We consider this feature to be very important since it allows us to know which type of work is better paid. However, due to its high percentage of missing data, it is impossible to impute it without having an impact on the result. Therefore, we decided to create a new category. | New category created: “Not considered” |
| 9 / 10 | major industry code / major occupation code | These two features are closely related (they have the same percentage of missing data), and provide us with information about the company where the individual works. | New category created: “Not considered” |
| 26 / 27 / 28 | migration code-change in msa / migration code-change in reg / migration code-move within reg | These three features are highly related (same percentage of missing data), and provide us with information on internal migrations within the US. We do not consider it a key factor for analysis, and since it has a very high percentage of missing data, we decided to remove these features. | Remove |
| 29 | live in this house 1 year ago | This feature provides us with information on whether the individual has a stable home. We do not consider it important. | Remove |
| 32 | family members under 18 | It provides us with information about the individual's family (whether they have minors in their family). We do not consider it important, as it may affect spending, but in this report, we are interested in income. | Remove |

We start the analysis with group 3: We will only attempt to impute those variables that are essential for the analysis of a person's salary, as this group has a very high percentage of missing data.

| **#** | **Columna** | **Analysis** | **Results** |
| --- | --- | --- | --- |
| 7 | enrolled in edu inst last wk (whether a person was enrolled in an educational institution (such as a school or university) during the last week) | We have seen that it can only take two values: "College or university" and "High school". This variable can be useful when analyzing, for example, aspects such as the person's free time, but we do not consider it a very important factor. | Remove |
| 14 | member of a labor union | It's a very interesting variable, we can see for example the salary difference between those who are members and those who are not. However, it is very difficult to impute in this case, as it is difficult to deduce whether a person has joined a union or not. | Remove |
| 15 | reason for unemployment | This variable only provides additional information about the people's position, we do not consider it important. | Remove |
| 21 | region of previous residence | Like the previous variable, it only provides us with information about their previous residence, so we do not consider it important. | Remove |
| 22 | state of previous residence | Como la variable anterior, es información sobre donde ha vivido anteriormente, no lo | Remove |
| 30 | migration prev res in sunbelt (whether the person had lived in a state that is part of the Sun Belt region of the United States in the previous year) | Like the previous variable, it provides information about where the person lived previously, but it is not considered important. | Remove |
| 38 | fill inc questionnaire for veteran's admin  (likely refers to whether the individual completed a questionnaire for the Veterans Administration in order to receive income-related benefits) | The "fill inc questionnaire for veteran's admin" variable is related to whether the respondent filled in a questionnaire for the Veteran's Administration (VA), which is an agency of the federal government that provides benefits and services to veterans and their families. The "veterans benefits'' variable refers to whether the respondent received any benefits from the VA. These two variables are related because filling in a questionnaire for the VA is often a step in the process of applying for veterans benefits.  Since the missing data % for VA is 0%, we can deduce this variable from VA. We could impute it, but we wouldn't gain any additional information, since we already have enough with the VA variable. | Remove |

# 

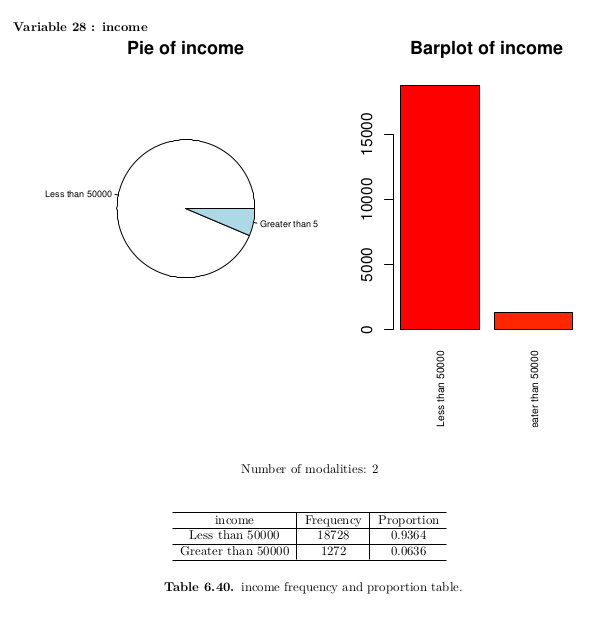
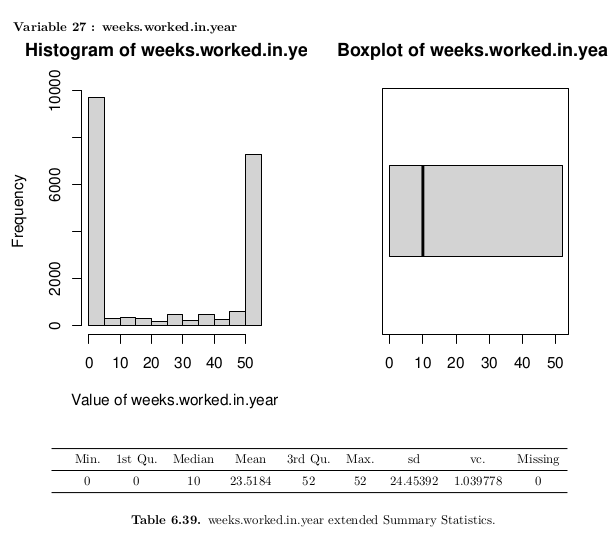
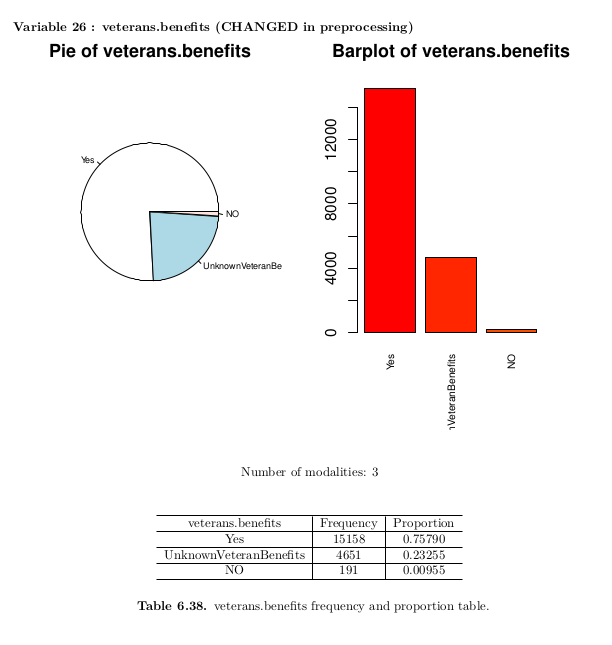
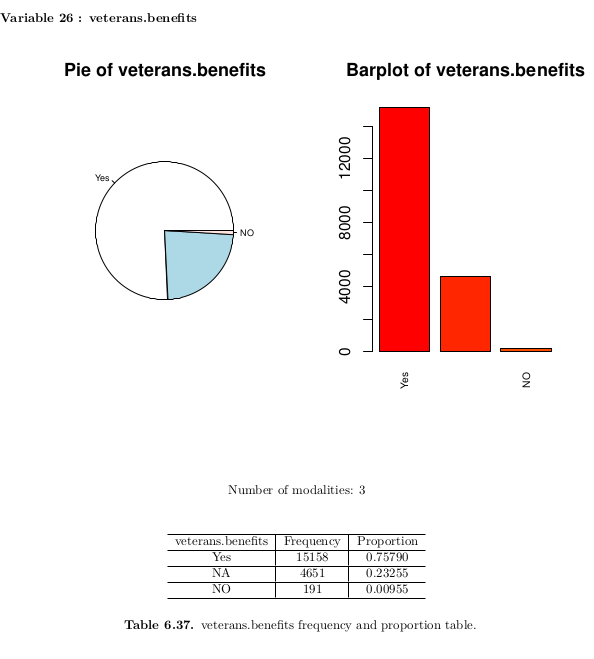
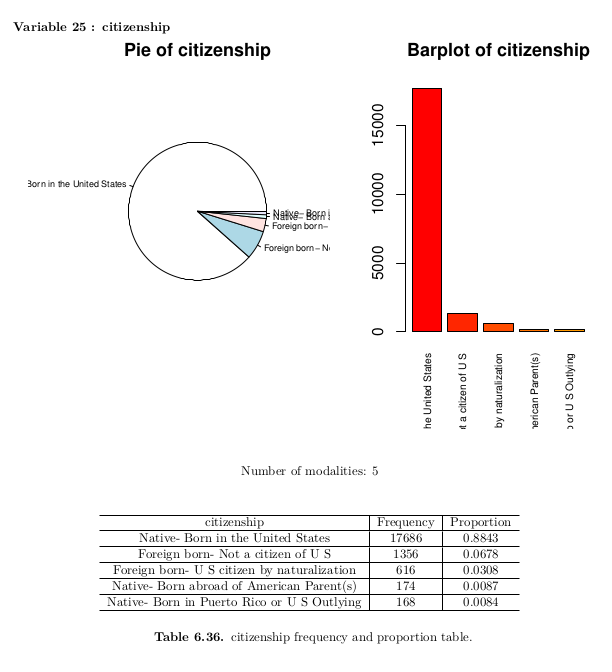
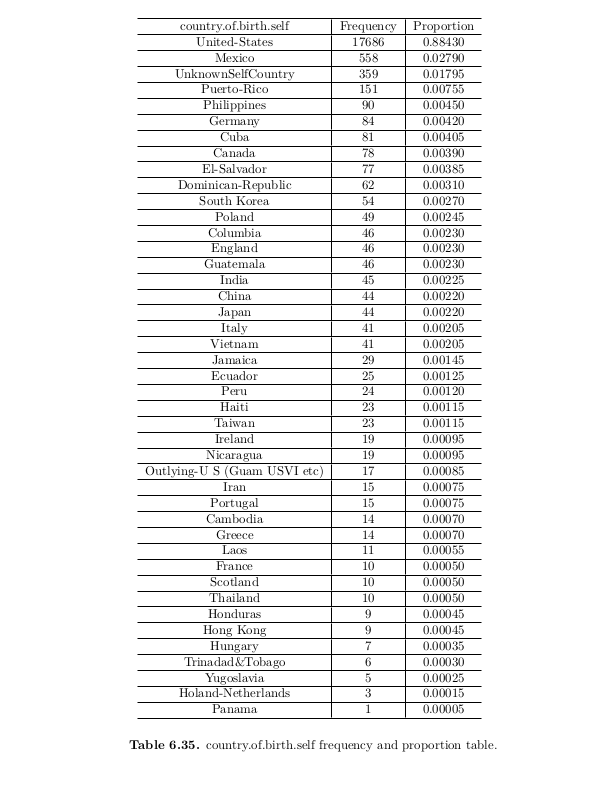
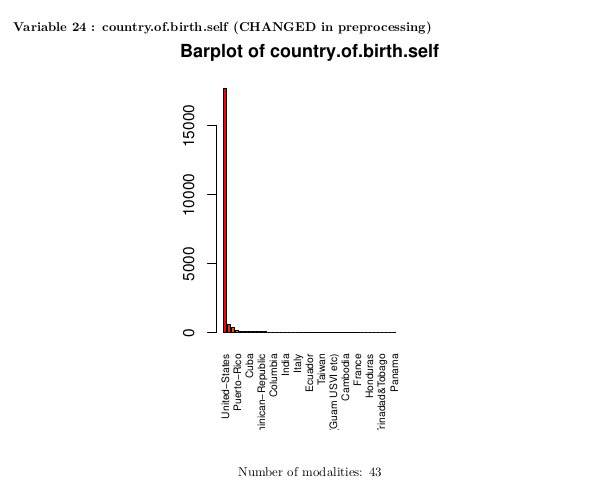
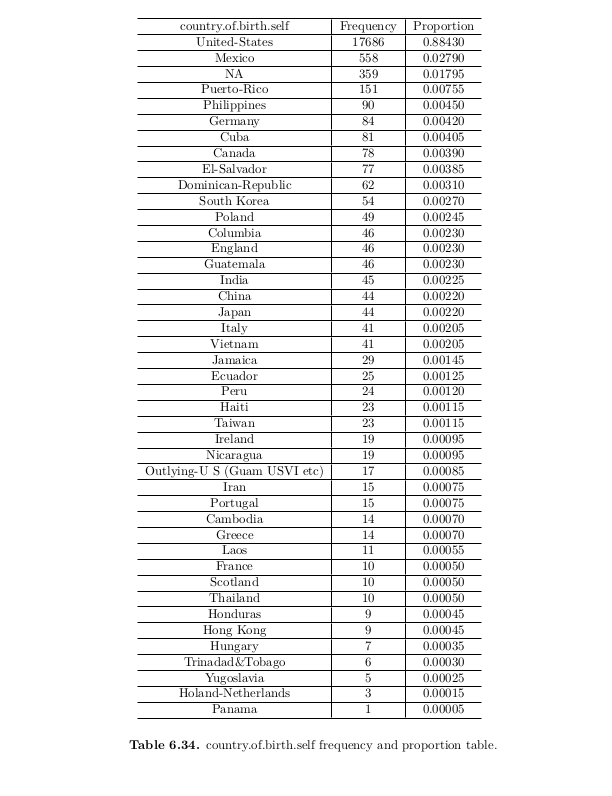
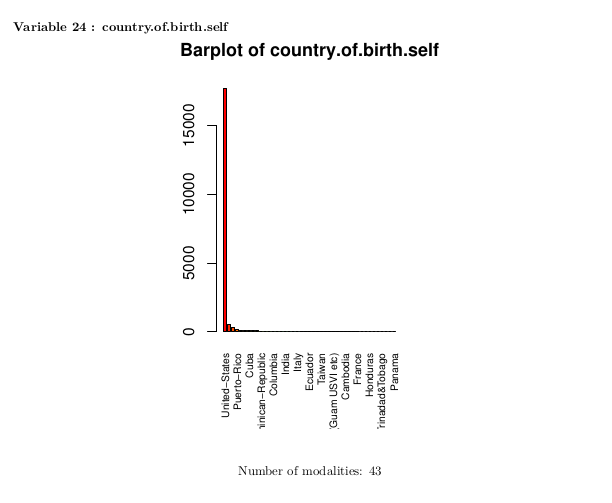
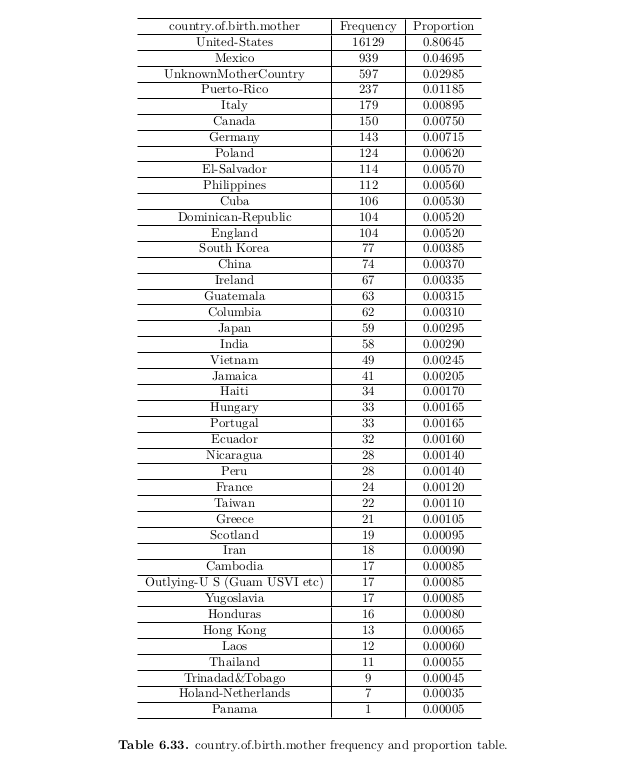
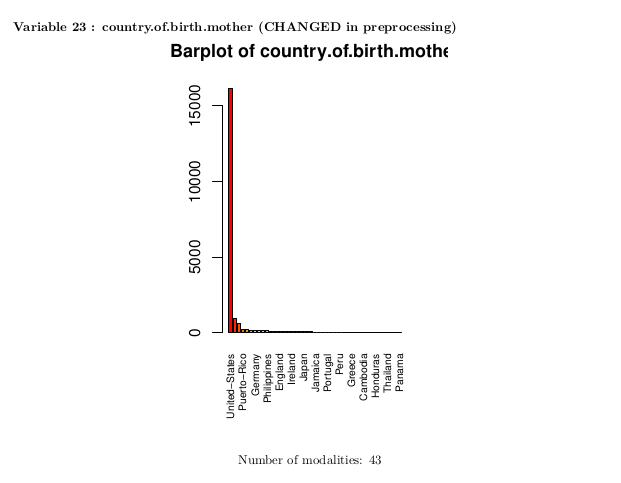
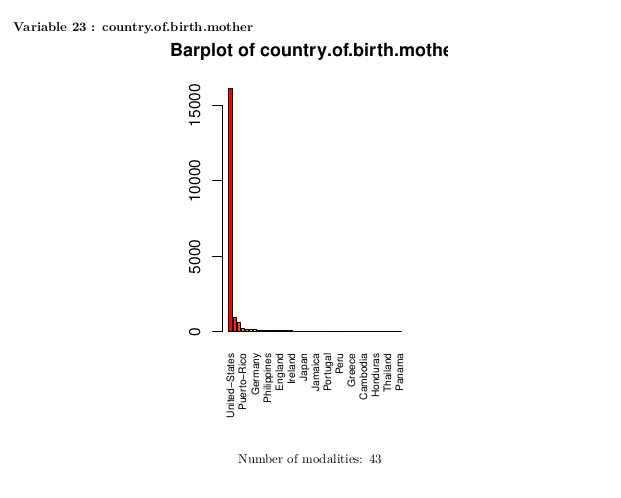
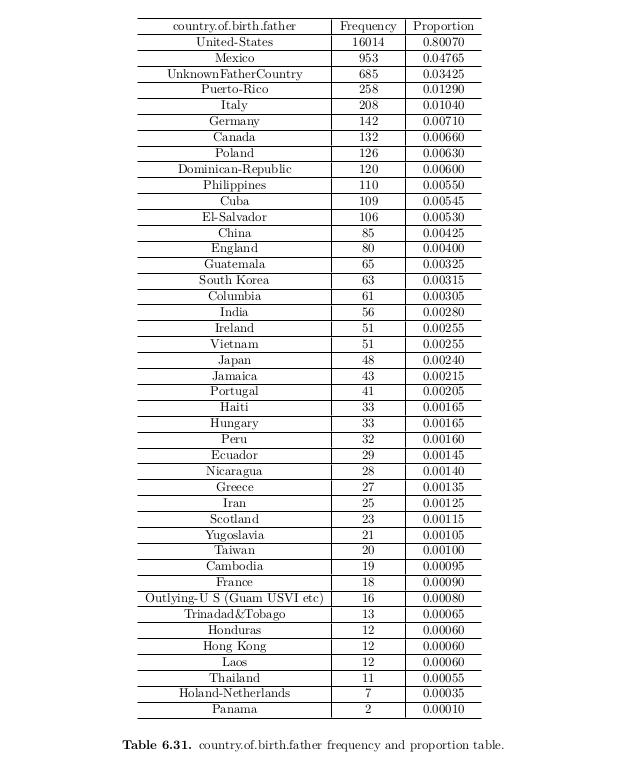
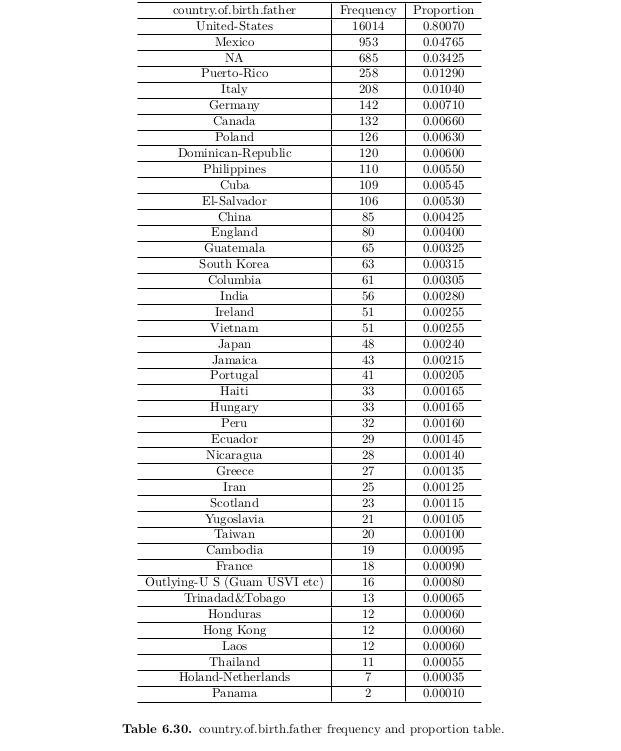
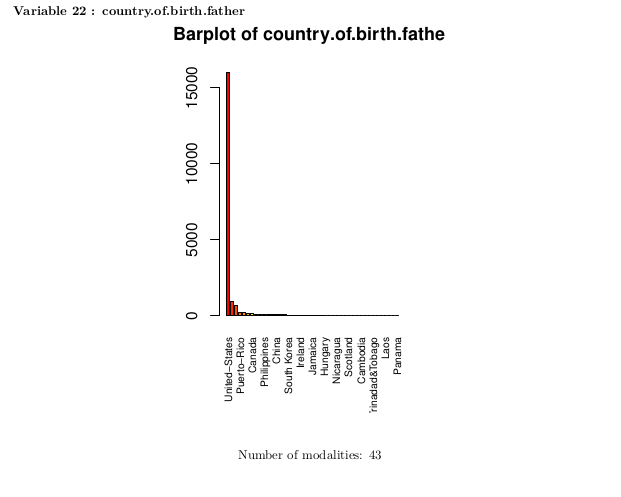
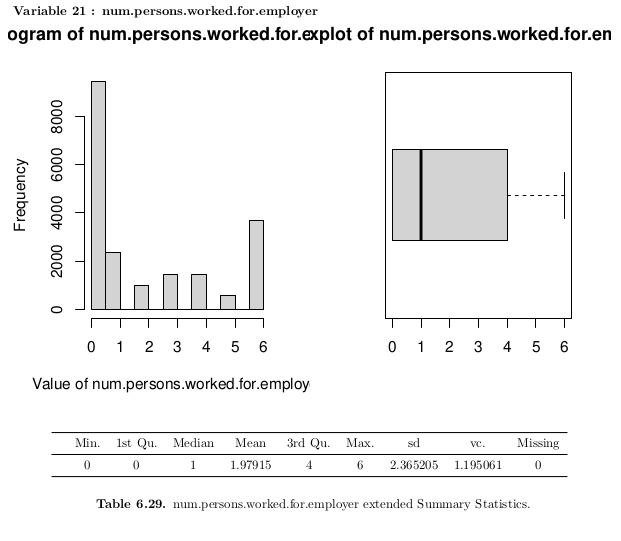
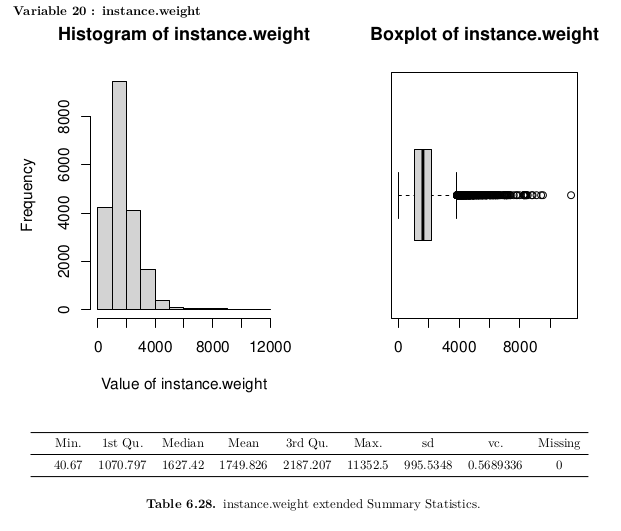
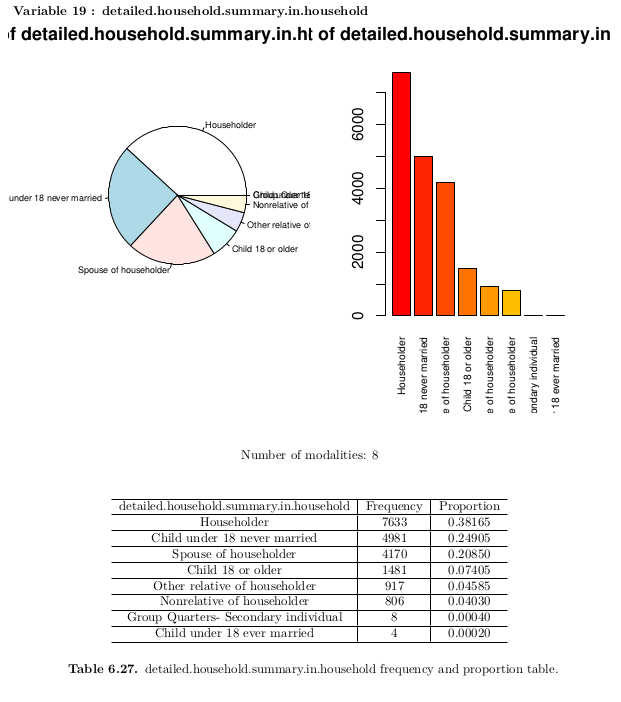
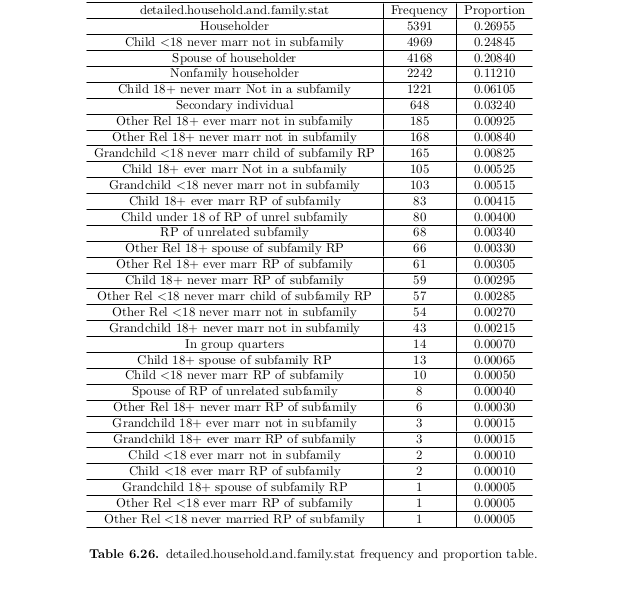
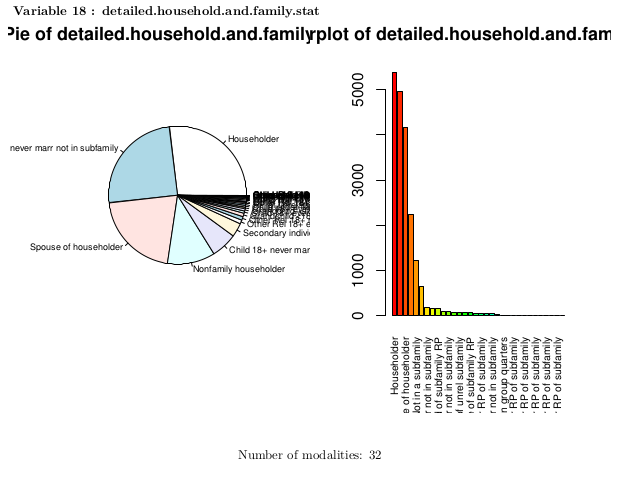
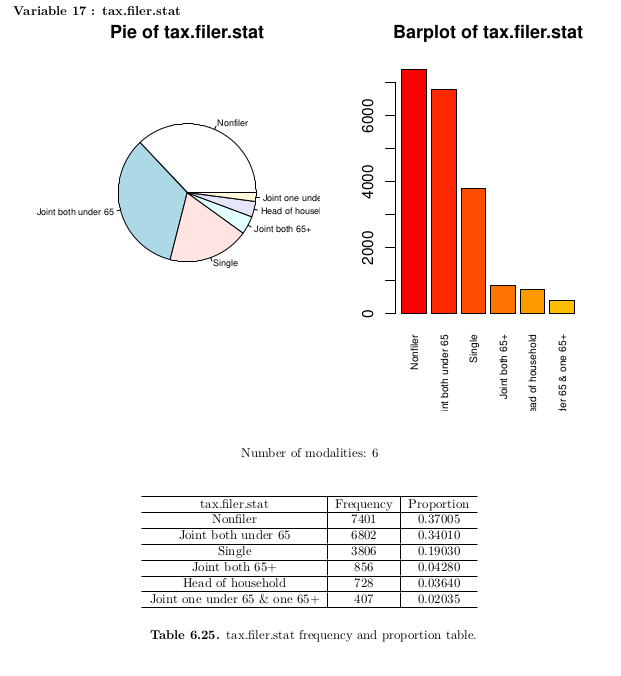
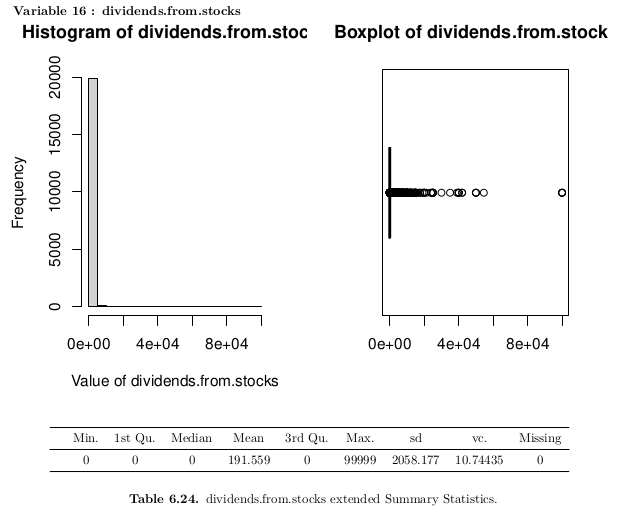
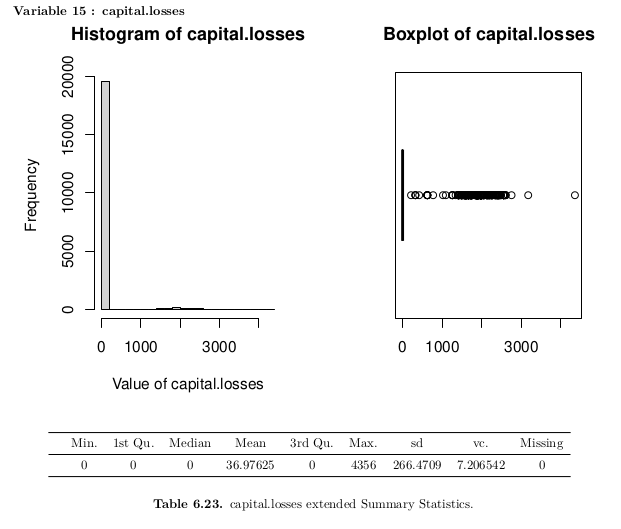
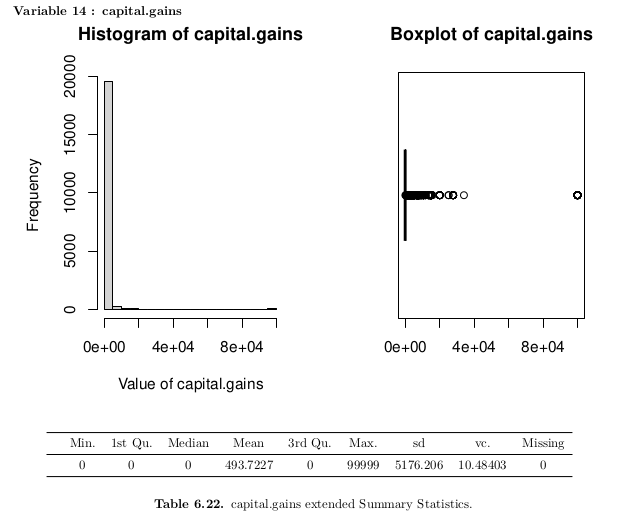
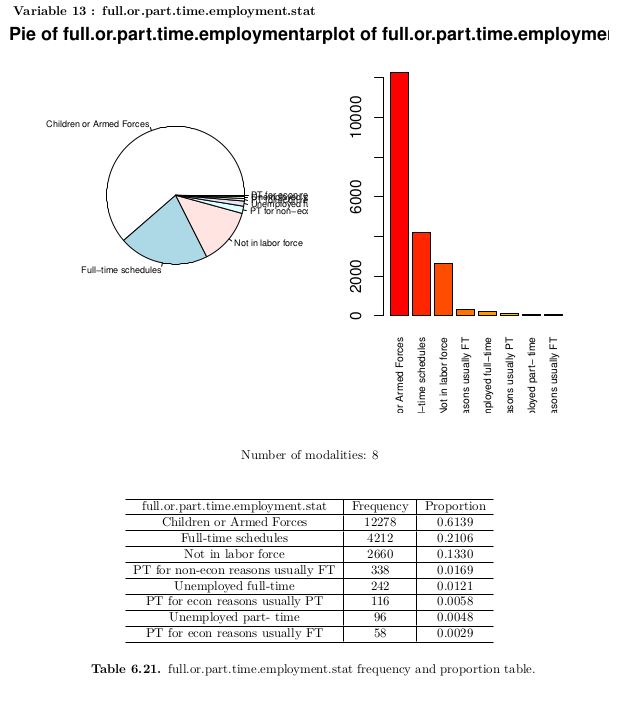
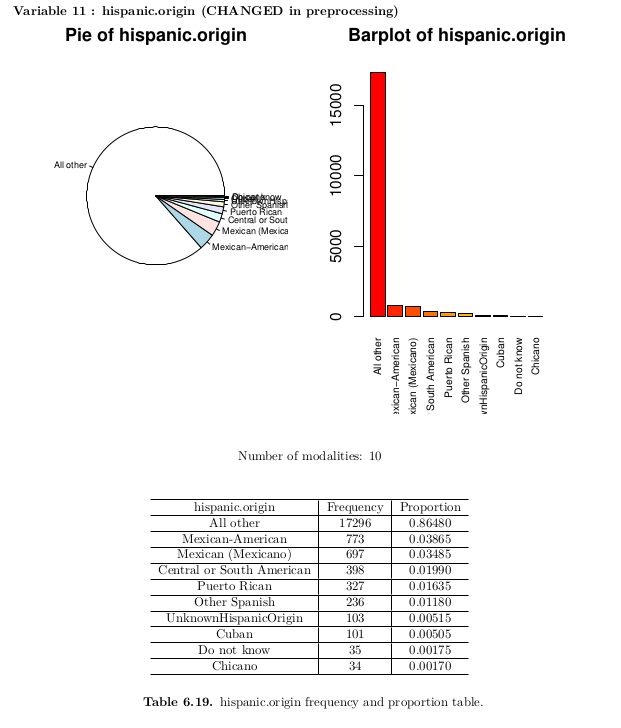
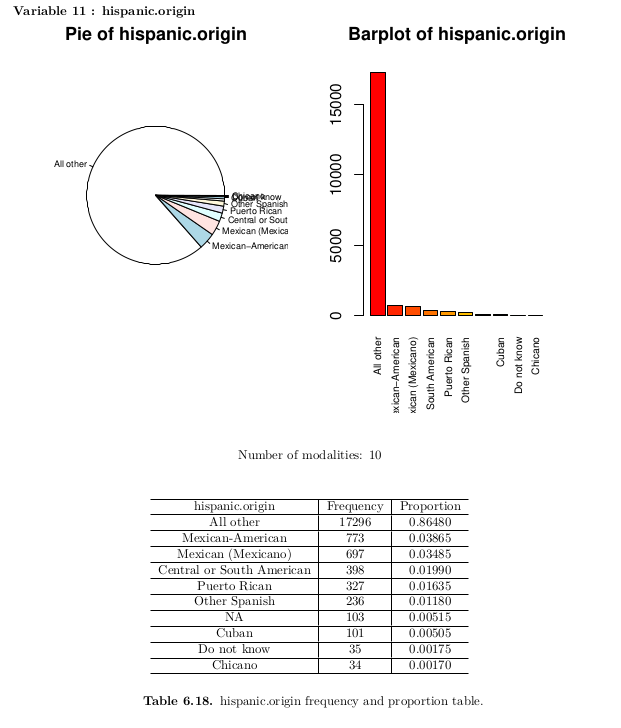
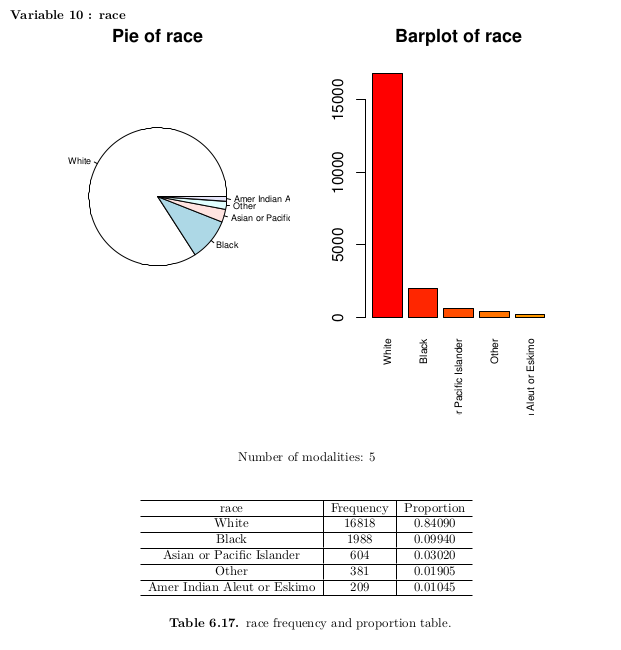
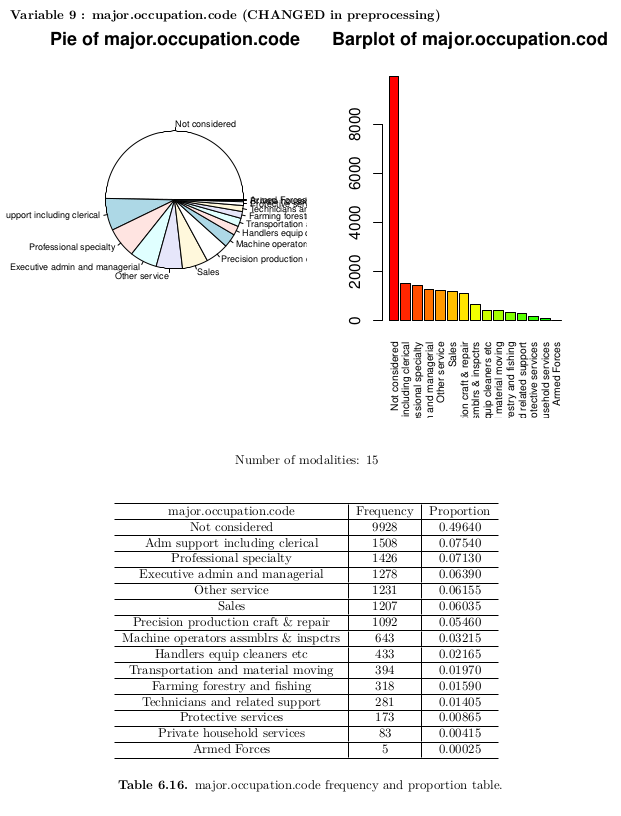
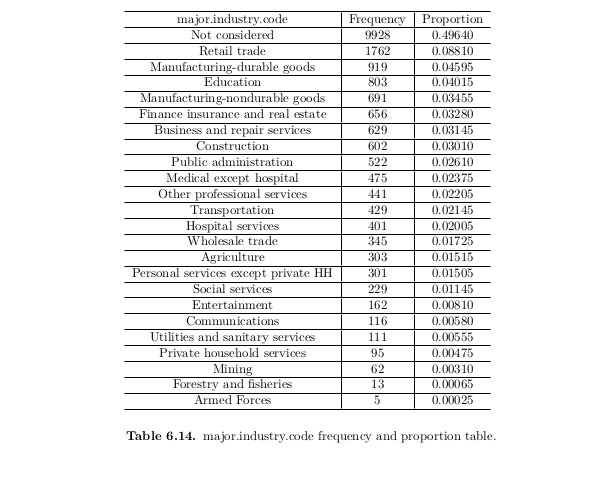
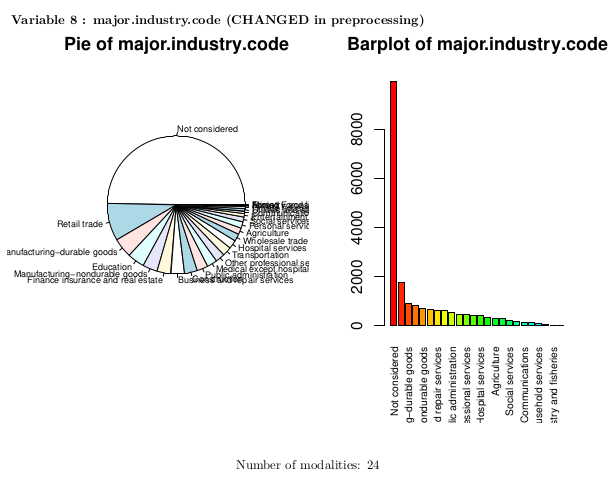
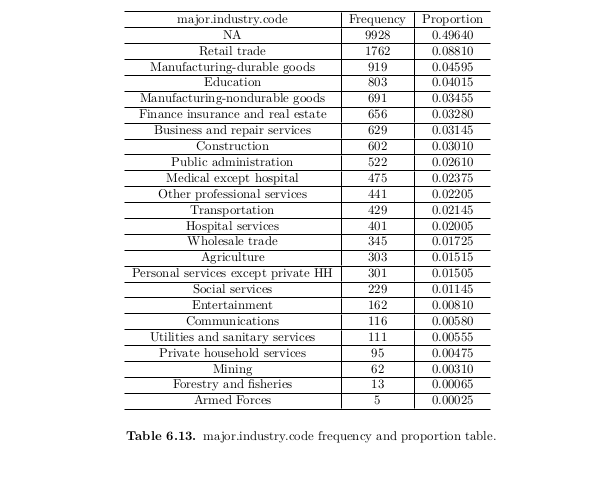
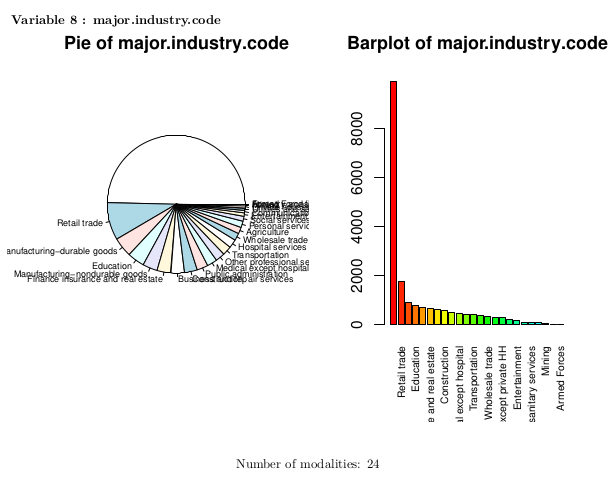
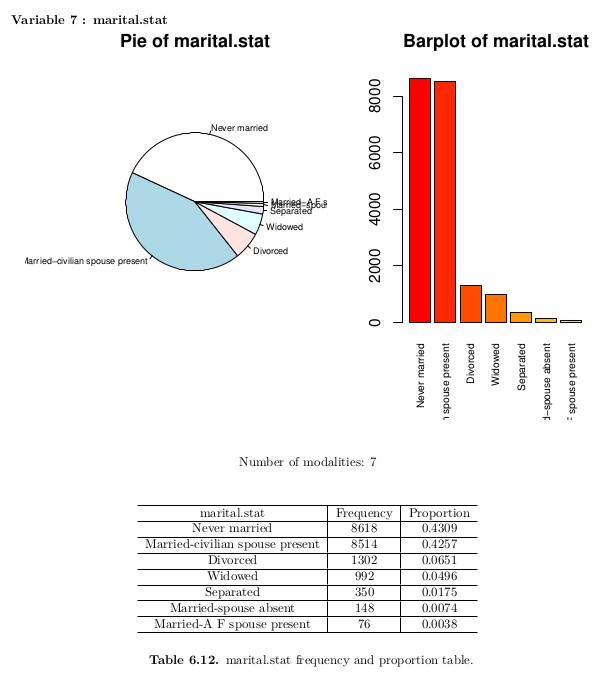
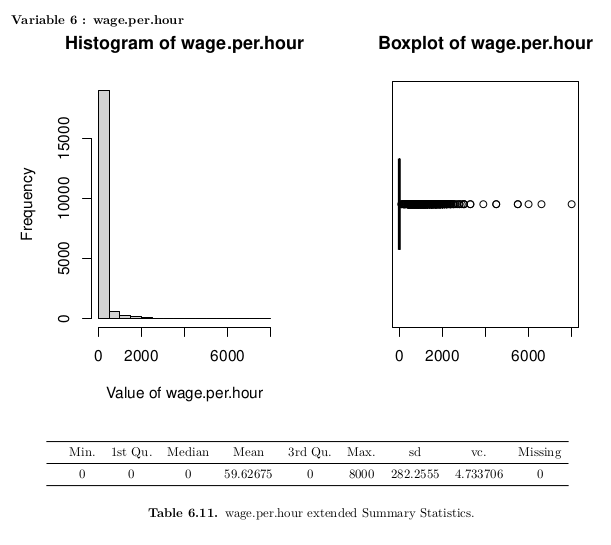
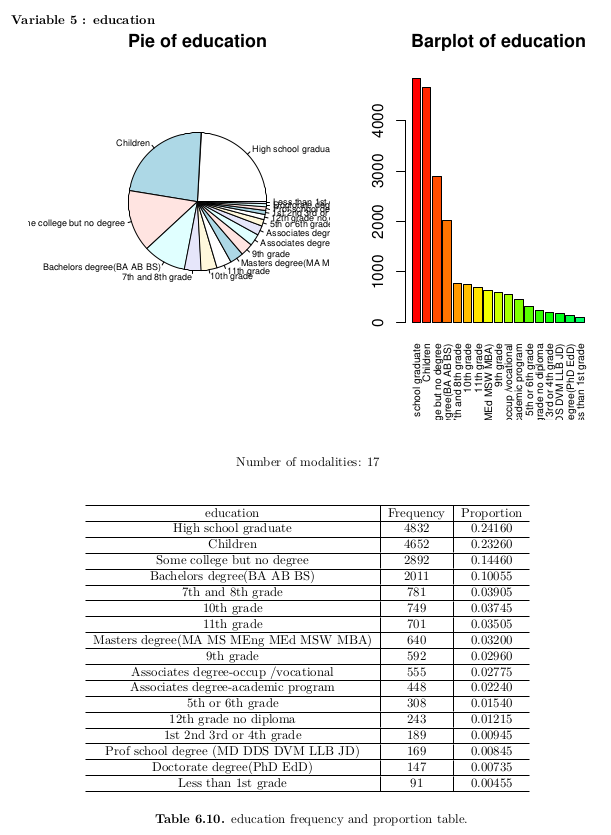
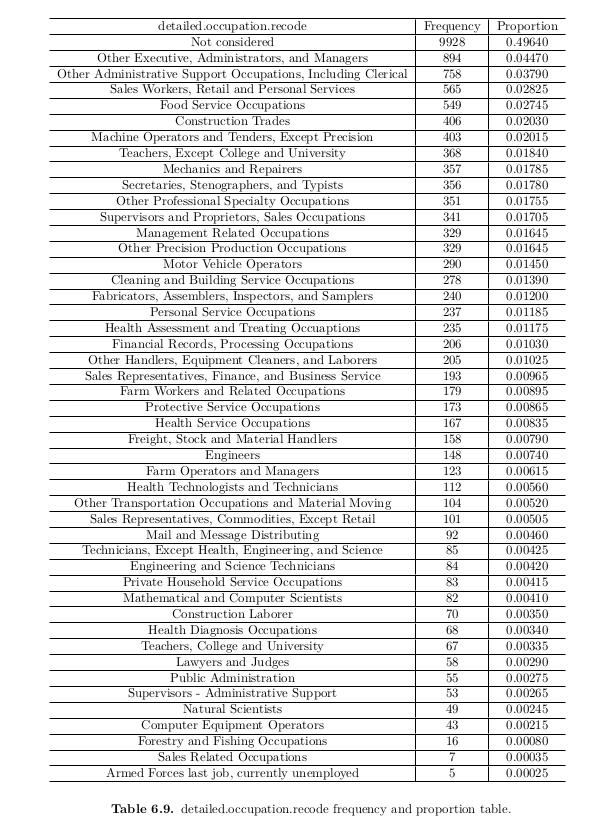
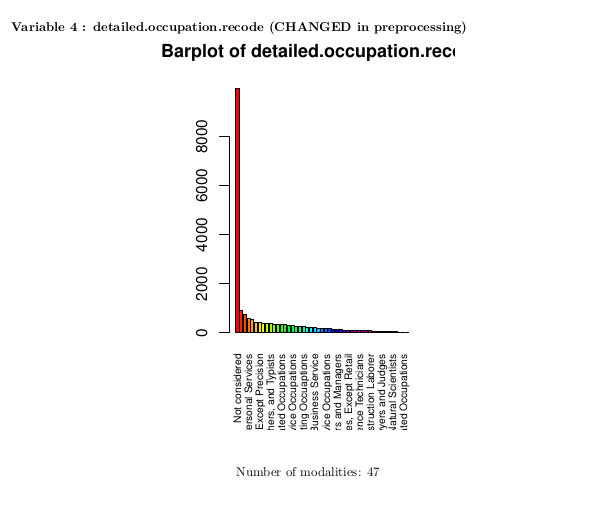
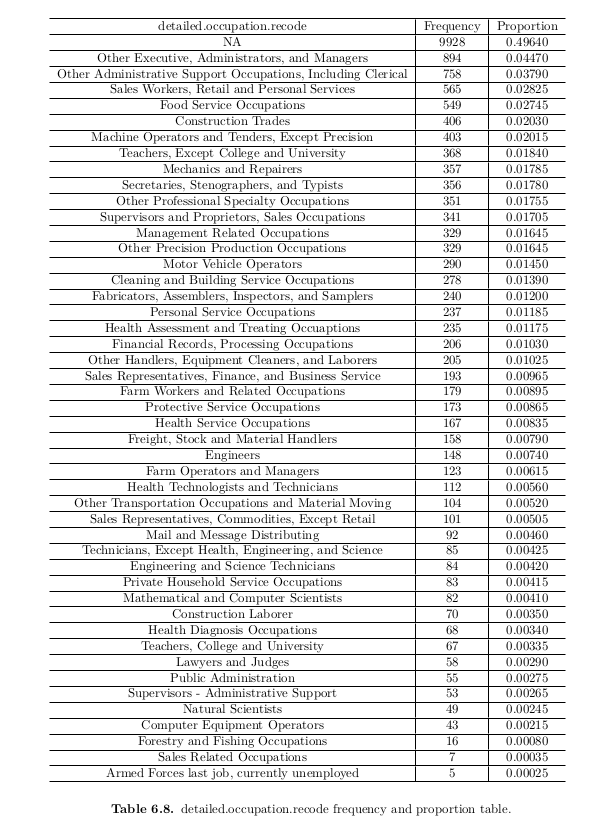
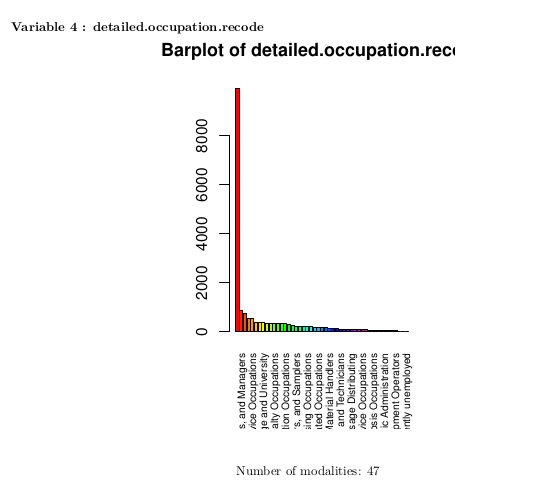
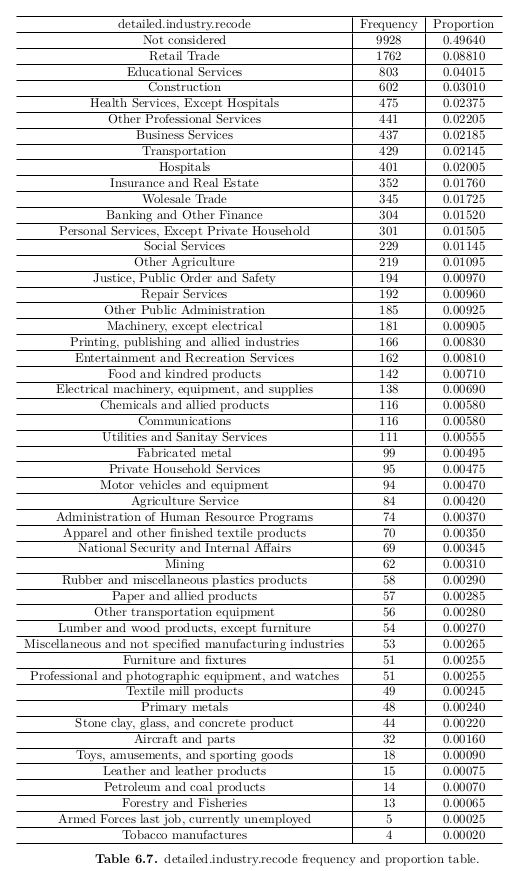
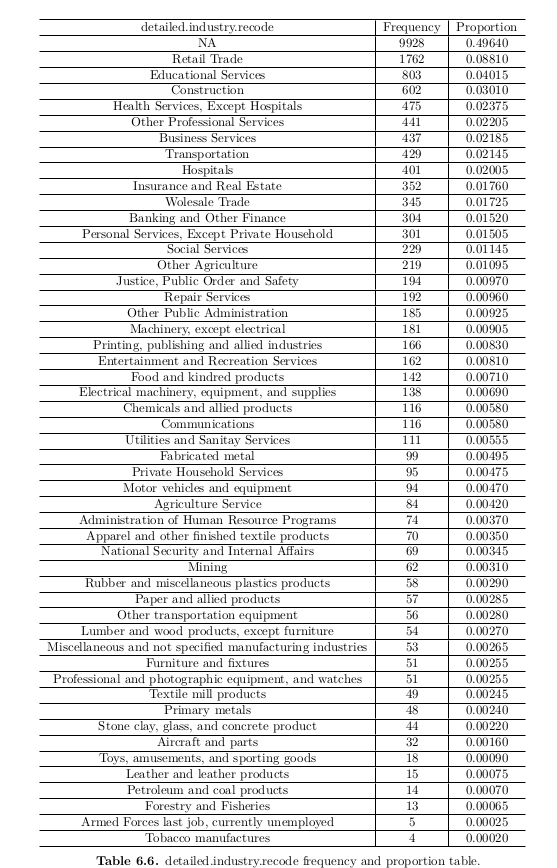
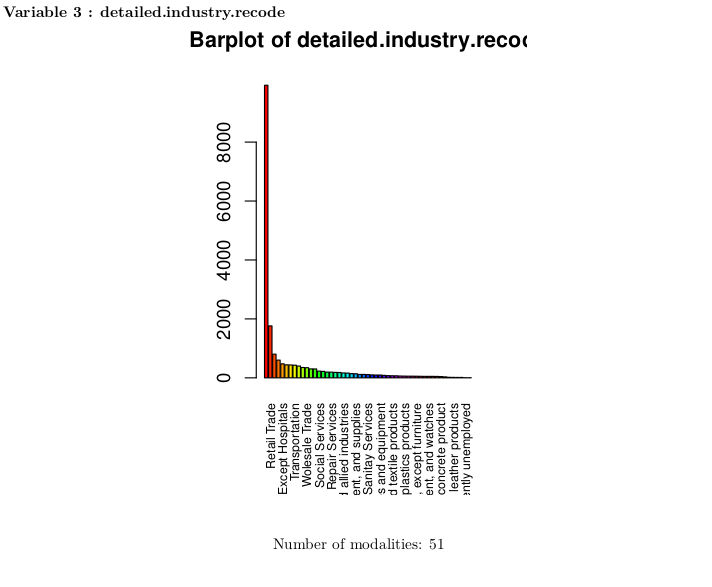
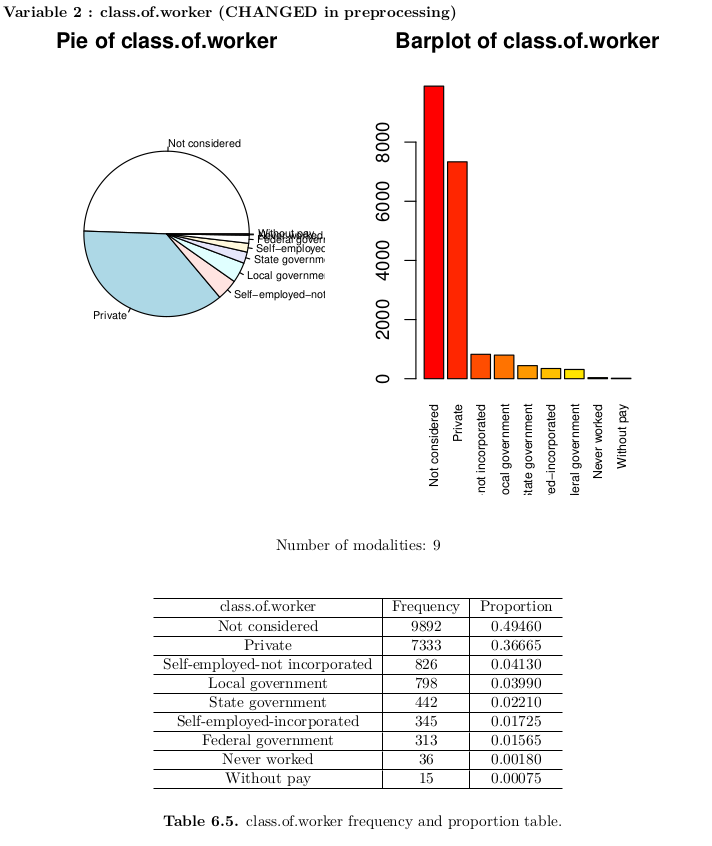
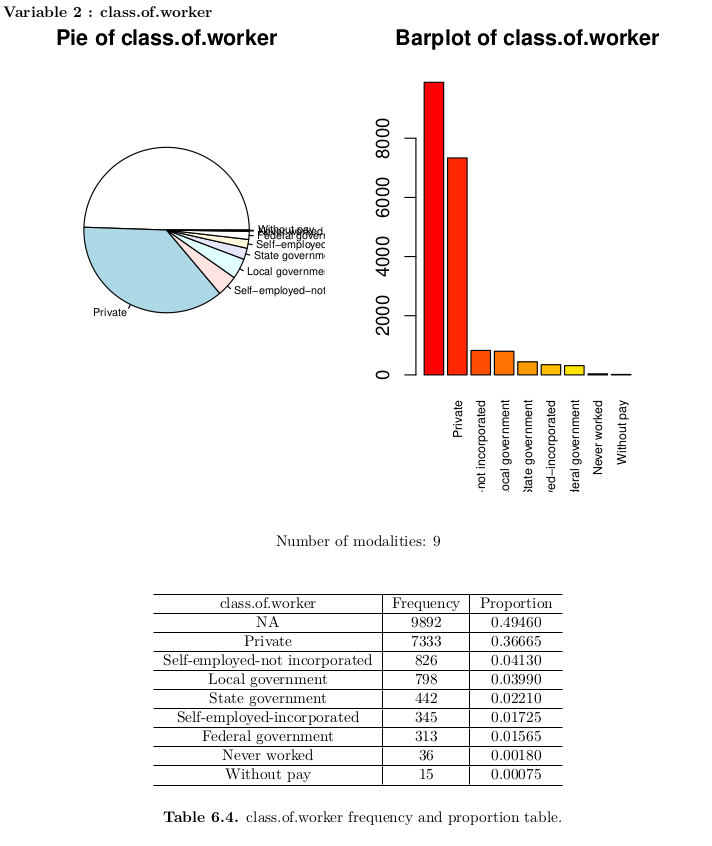
# 4.Basic statistical descriptive analysis

To obtain the fundamental univariate statistics, an RMarkdown script was utilized to automatically generate descriptive visualizations and tabulations for each variable within the income dataset. Furthermore, if any variable was impacted during the preprocessing imputation phase, visualizations are displayed both pre- and post-modification. Different information is generated depending on the two types of our variables:

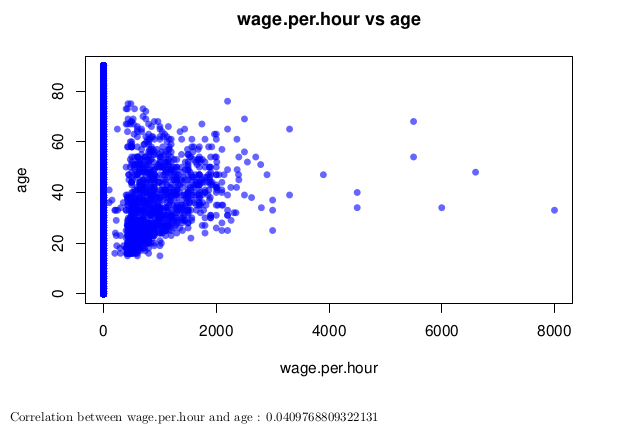
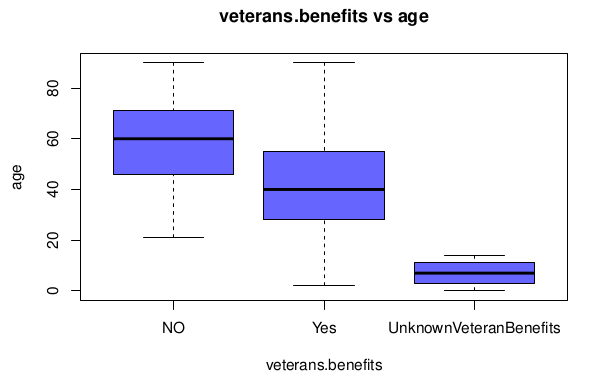
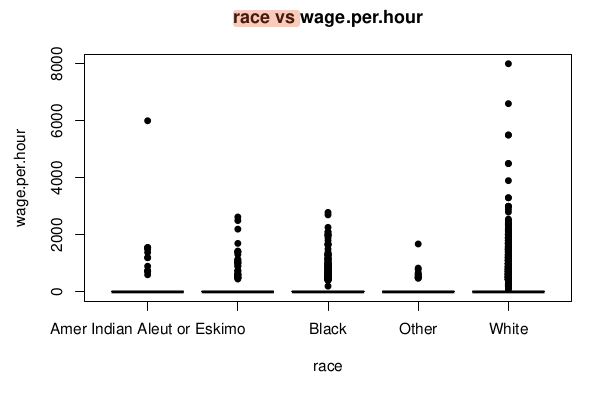
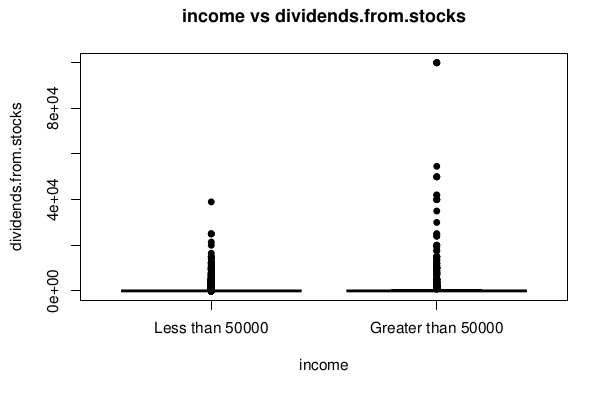
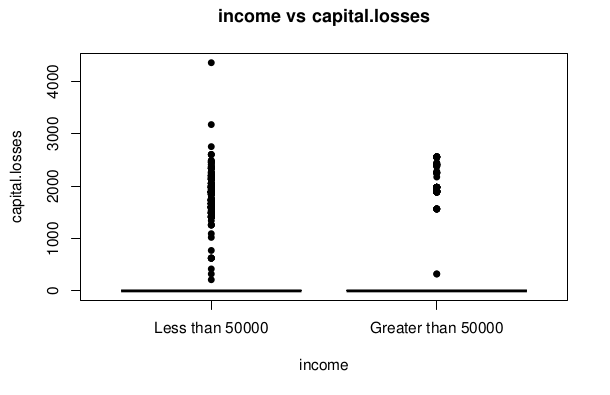
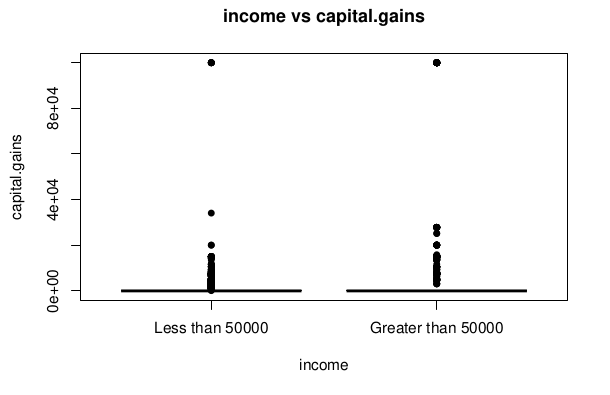
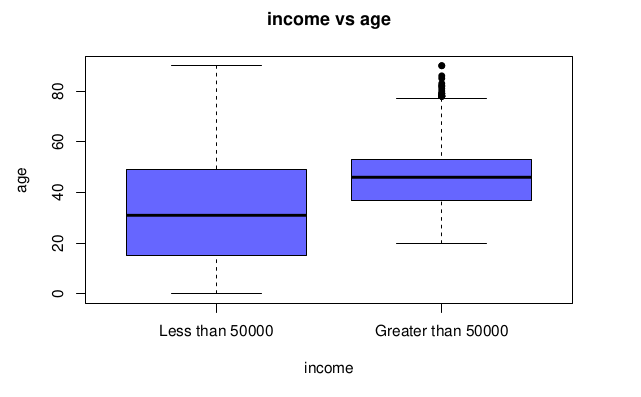
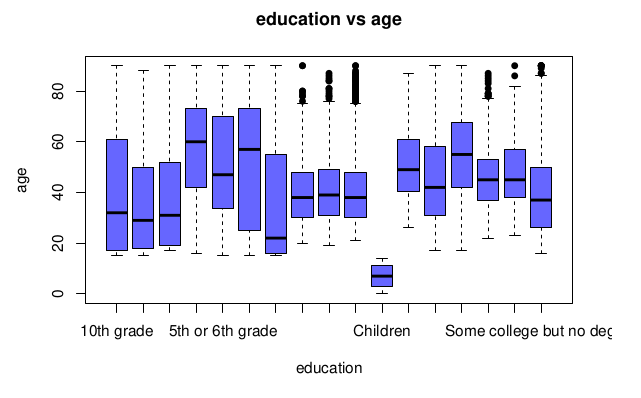
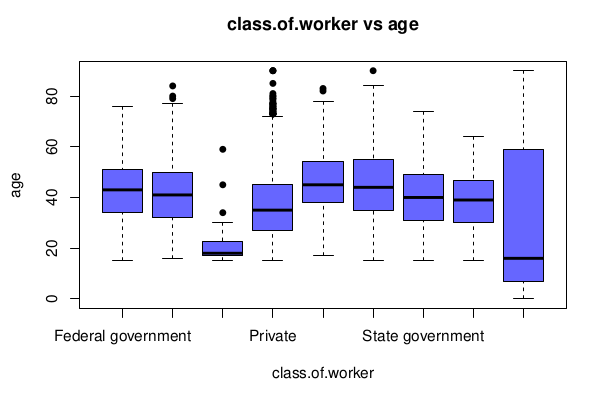
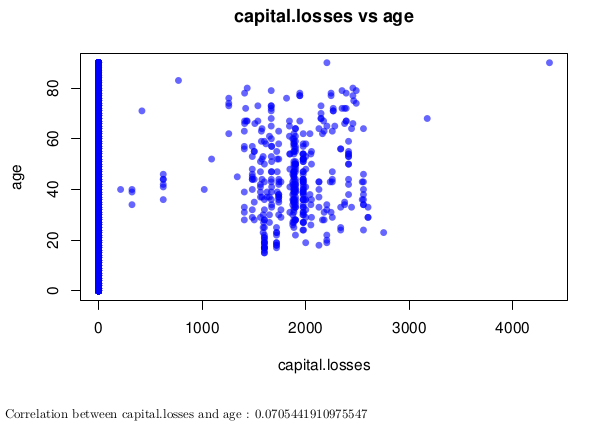
* For categorical variables, a pie chart and bar plot were computed. Only legible pie charts were included. Subsequently, the quantity of distinct modalities is displayed, followed by a tabulation containing the count for each modality and the relative frequency of each factor expressed as a proportion.
* For numerical variables, we generated a histogram and a box plot. Additionally, a tabulation displays the minimum and maximum values for the variable in question, along with the 1st, 2nd (Median), and 3rd quartiles; as well as the mean, standard deviation, coefficient of variation and quantity of unknown values.

With respect to bivariate statistics, the RMarkdown script was also employed to generate visualizations contrasting numerical variables with all other variables within our dataset. Following generation, a selection of visualizations deemed to provide additional information to the analysis were chosen. Additionally, for numerical variables, we computed the correlation with the aforementioned variables.

## 4.1 Univariate analysis



## 4.2 Bivariate plots



## 4.3 Conclusion

From the statistical descriptive analysis conducted on our income dataset, several conclusions can be drawn. Upon examination of the histograms and boxplots for the numerical variables, it can be observed that there is a relative variability in the age variable, the majority of data points are concentrated between 20-50 years of age. Additionally, two other numerical variables exhibit high variability: *num.persons.worked.for.employer* and *weeks.worked.in.year*. The other ones have a low variability.

In terms of categorical variables, there is a greater degree of variability present. However, it should be noted that in some instances one category may predominate over others by more than 60%. For certain variables such as *industry recodes*, *occupation recodes*, *major industry code*, *major occupation code* and *class of worker code*; the predominant category is “Not considered” due to the inclusion of a large number of children or retired persons in the dataset.

Upon conducting a bivariate analysis of our dataset, it can be observed that there is a significant degree of variability with respect to the variables of *capital gains*, *capital losses* and *wage per hour* in relation to *income*, *age*, *sex* and *education*. This suggests that these variables may have a considerable impact on the distribution of income across different age groups, between sexes, between different levels of education and also between different classes of workers.

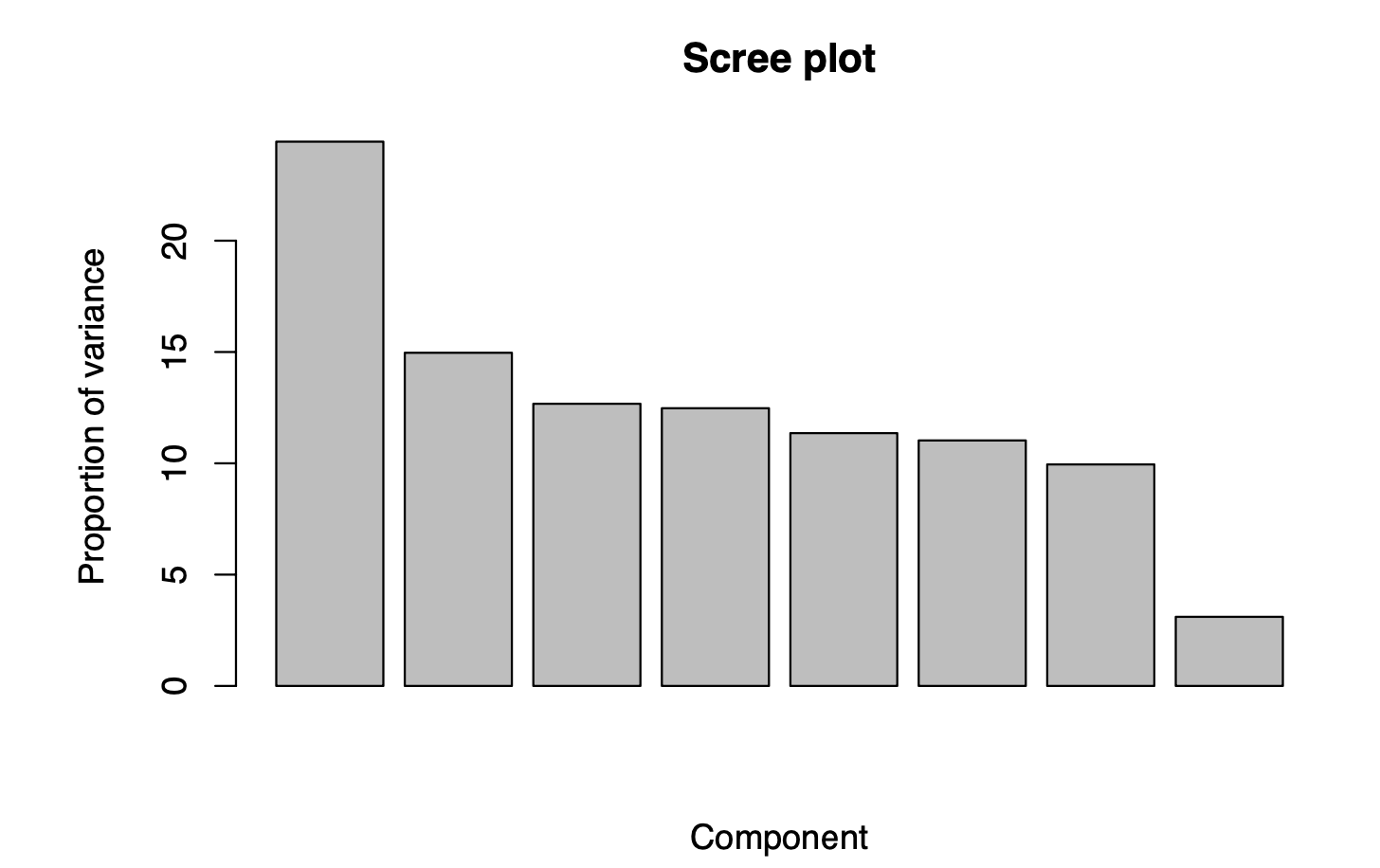
Finally, some general additional conclusions are:

* Firstly, it appears that individuals with an income greater than $50,000 exhibit less variability with respect to age compared to those with an income lower than $50,000. Additionally, those with an income greater than $50,000 have fewer capital losses and more dividends from stocks compared to those with an income lower than $50,000.
* The correlation between wage per hour and age is relatively low at 0.040. In terms of race, individuals who identify as white have a higher wage per hour compared to other races; however some individuals who identify as Indian also have a high wage per hour compared to other non-white races.
* In terms of class of worker and occupation, 36% are classified as private workers and 8% work in retail with the majority holding executive or managerial positions. The majority of the population (84%) identifies as white and 88% are US citizens. In terms of marital status, 42% are married while 43% have never been married.
* It is also worth noting that a significant proportion (61%) are either in the armed forces or children and 21% have a full-time schedule. Finally, the majority (93%) have an income less than $50,000 while only 7% have an income greater than $50,000.

# 5.PCA analysis for numerical variables

PCA is a useful technique for identifying patterns and relationships among variables and reducing the dimensionality of a dataset

## 5.1 Scree plot

The scree plot is a graph of the eigenvalues of the principal components, ordered by magnitude. The scree plot allows us to visualize the proportion of variance explained by each component and identify the number of components to retain.

In this graph, we can see that the principal components more or less explain the same amount of variance throughout the graph. To more directly see how many PCs we need to select we will take a look at the cumulative scree plot which will make it much easier to identify.



Here we can clearly see that pc number 5 gets really close to that 80%, so we have chosen those 5 components to represent our entire data. Ideally, we would only have 2 or 3 principal components to represent our data.

## 5.2 Factorial map

A factorial map is a graphical representation of the relationship between several variables in a dataset. It is a type of dimensionality reduction technique that is used to explore and visualize high-dimensional data. Factorial maps can be created using methods such as principal component analysis (PCA).

In a factorial map, each data point (e.g., a sample or observation) is represented as a point in a low-dimensional space (e.g., two-dimensional or three-dimensional space), where the distances between points reflect the similarities or dissimilarities between the data points. The position of each point in the map is determined by its scores on the underlying factors or dimensions that explain the most variation in the dataset.

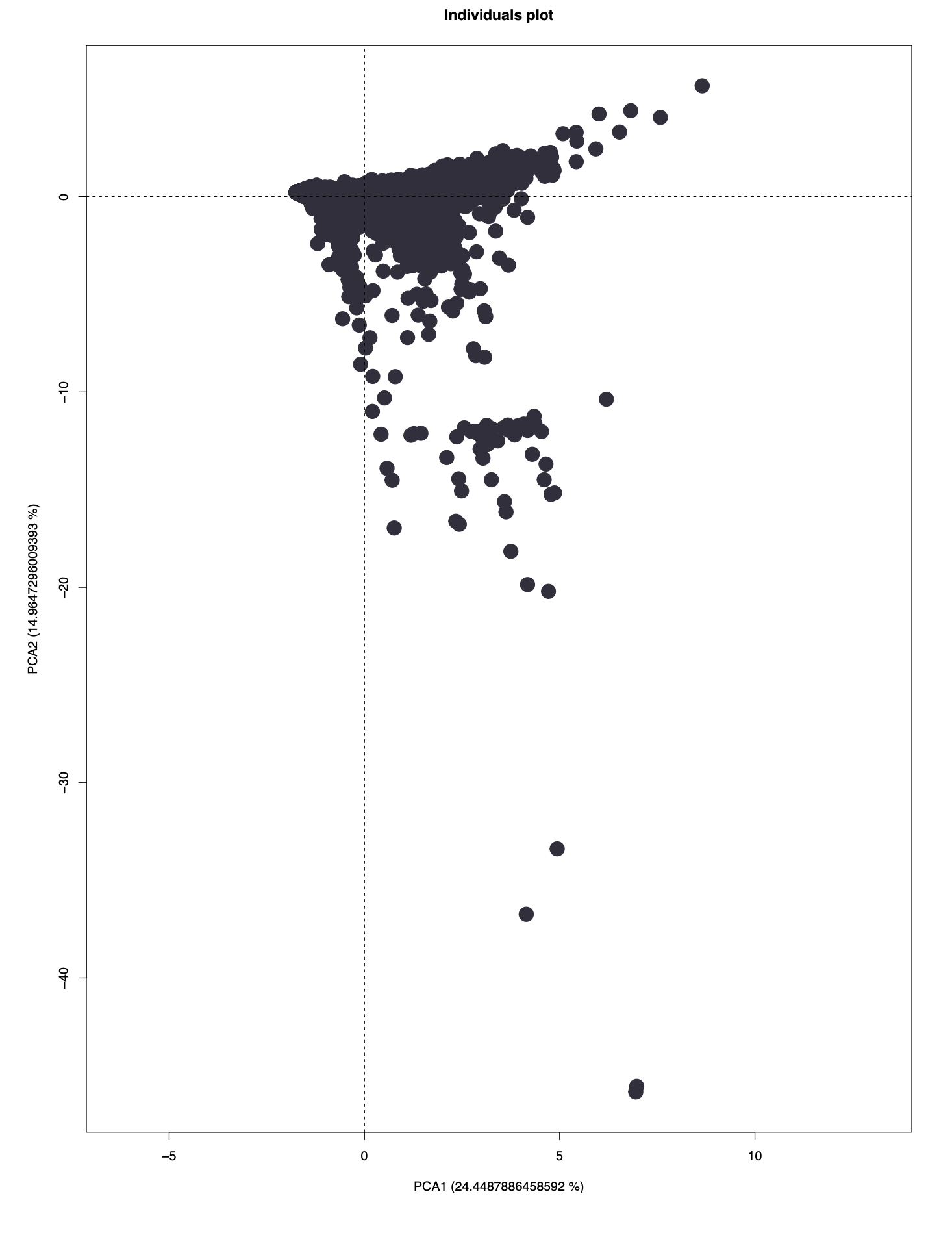
Factorial maps are useful for gaining insights into complex datasets, identifying patterns and trends, and detecting outliers or anomalies. They can also be used for exploratory data analysis, clustering, classification, and visualization of high-dimensional data.

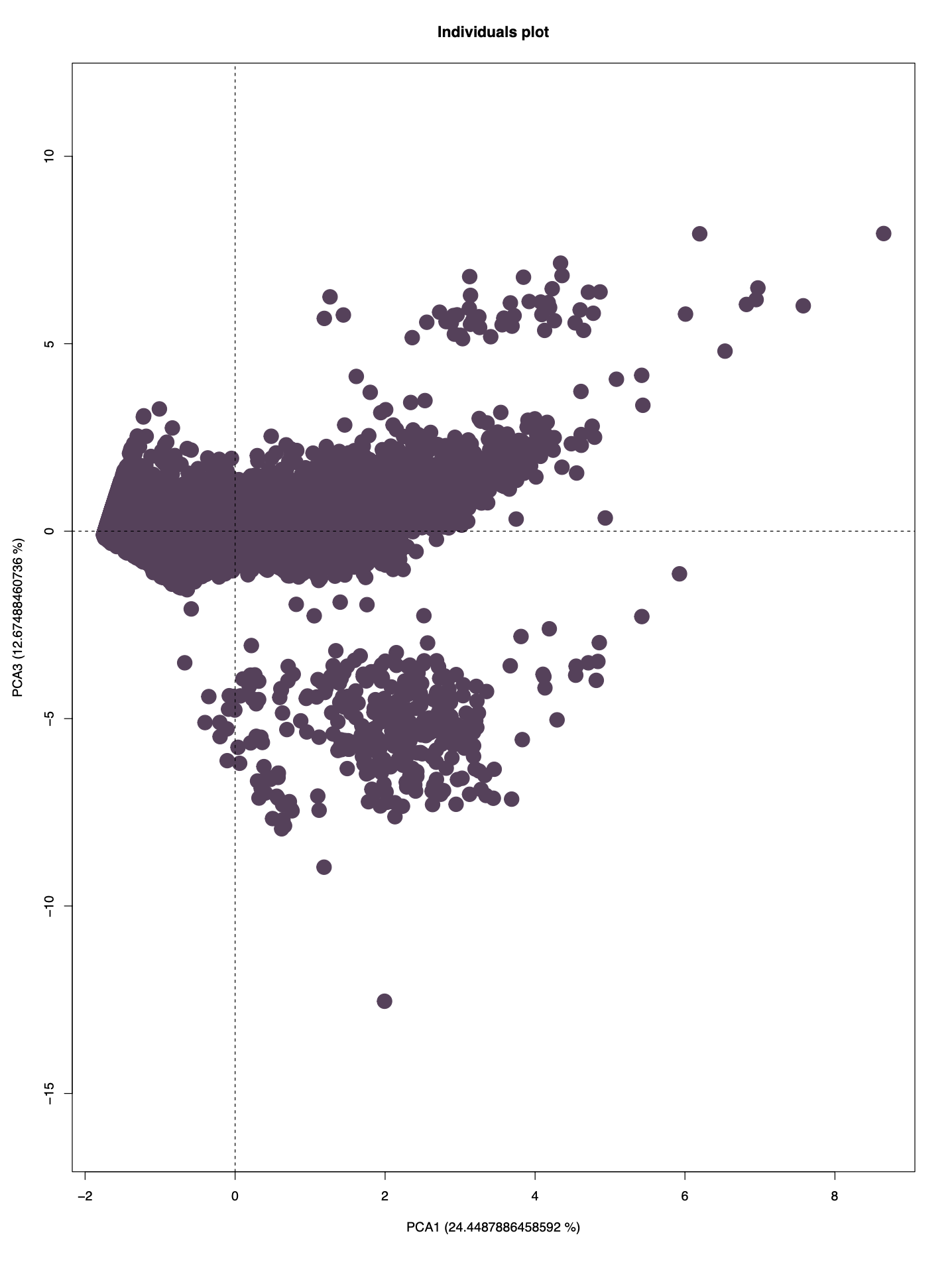
### 

### 5.2.1 Individuals plot

The individual's plot is a visualization of the observations in the two-dimensional space. Each observation is represented as a point, and the proximity of the points indicates similarity in the underlying variables.

In this plot, we can see individuals plotted on the two principal components which have the most variance. We can see that most of the individuals are close to the centre of the plot, but there are some outliers and even clusters of individuals which are further down the principal component 2.

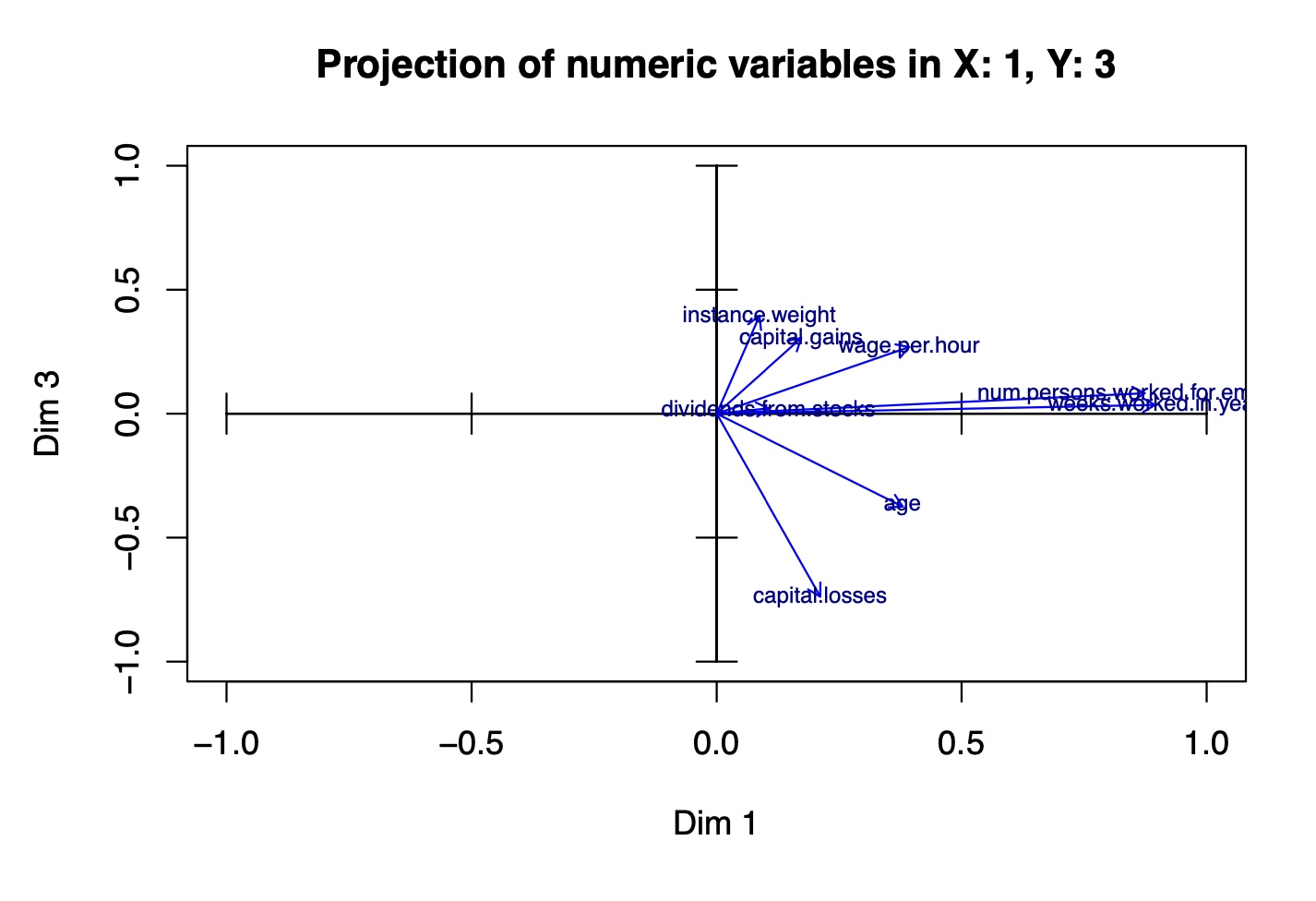


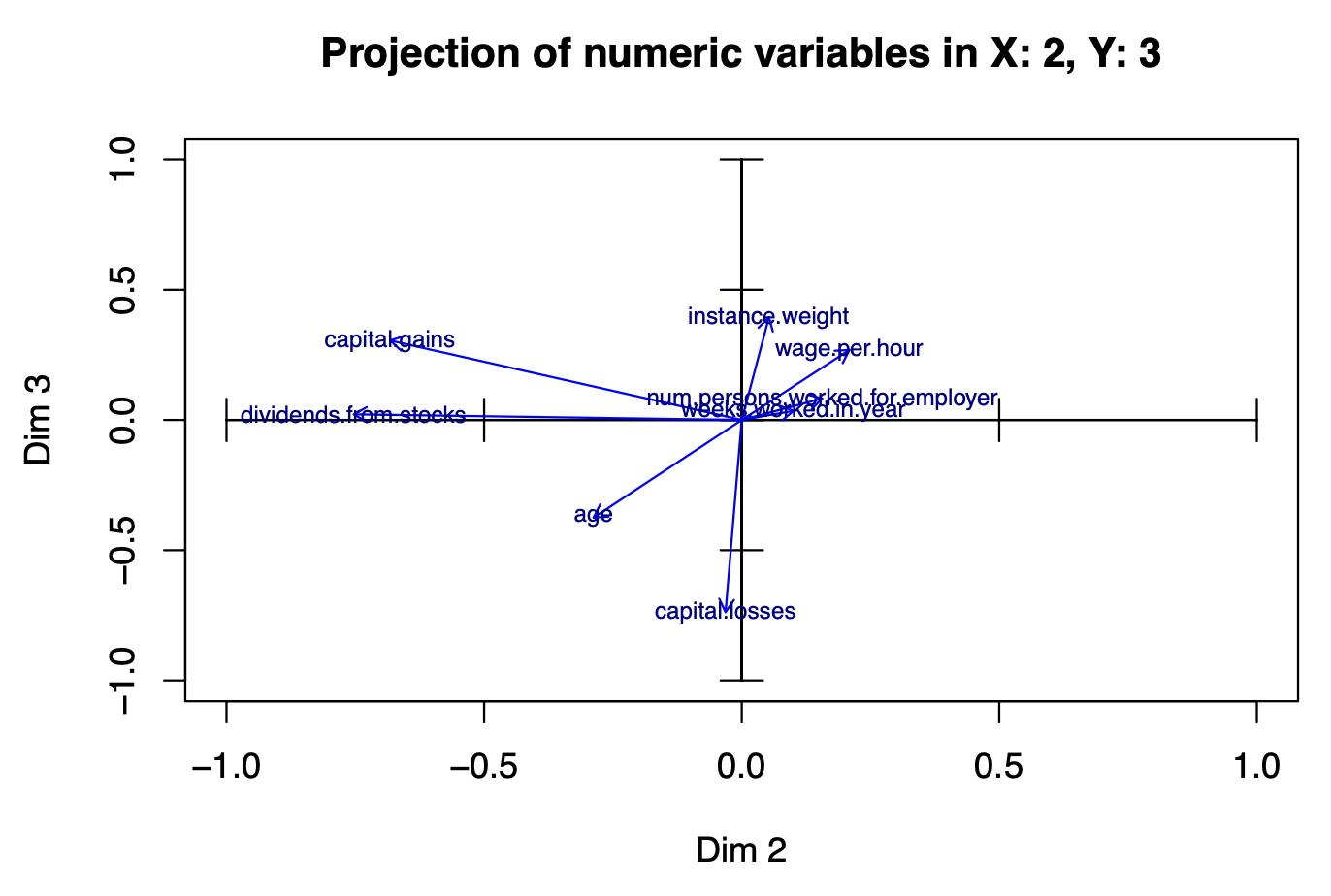


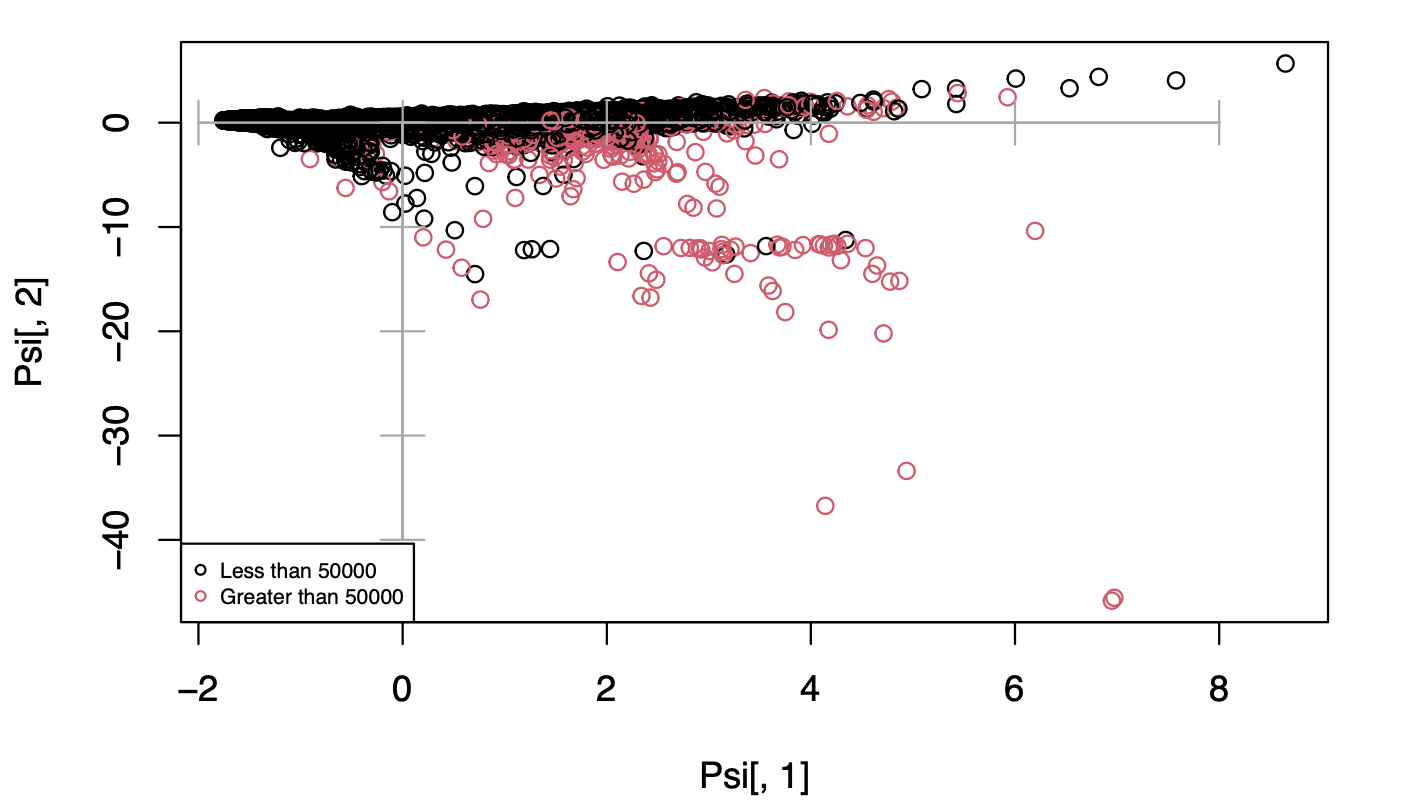
### 5.2.2 Common projection of numerical variables and modalities

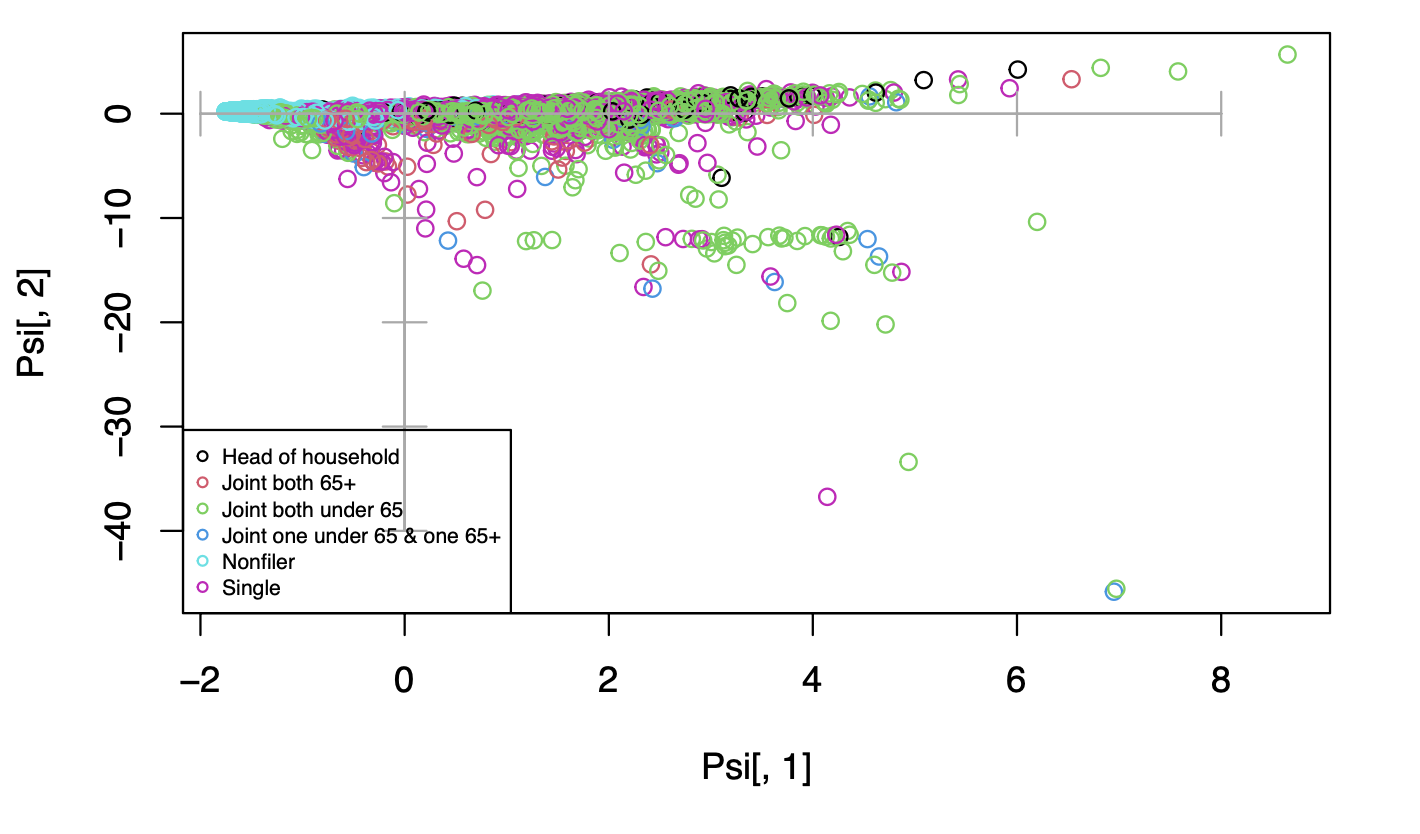
The common projection of numerical variables and modalities is a visualization of the variables in the two-dimensional space. Each variable is represented as an arrow pointing in the direction of the most important component, and the length of the arrow indicates the importance of the variable in that component.



From this plot, we can see that the most important variables on the y-axis are dividends from stocks and capital gains. In terms of the x-axis, the most important are the weeks worked in a year and the number of persons worked for the employer.





****

### 5.2.3 Interpretation of relationships among the observed variables

Based on the factorial map, we can draw several conclusions about the relationship between the variables, for example capital gains, dividends from stocks and age have a positive correlation between them.

Wage per hour and age seem to be negatively correlated between each other in the y-axis.

Num persons worked for employer and weeks worked for a year also seem to be positively correlated between each other.

Overall, it seems that all numerical variables contribute positively to the x-axis on principal component 1.

### 5.2.4 Conclusion

These are the most important conclusions that we can extract from the data:

* 5 principal components were chosen to represent the data, although 2 or 3 components would have been preferred.
* Most observations were clustered around the center of the plot, with some outliers.
* Dividends from stocks and capital gains were the most important variables in the y-axis, while weeks worked in a year and num persons worked for employer were most important in the x-axis.
* Capital gains, dividends from stocks, and age were positively correlated, while wage per hour and age were negatively correlated in the y-axis. Num persons worked for employer and weeks worked for a year were positively correlated.

# 6.Clustering

With hierarchical clustering they can find the most probable clustering number. To do this, all the numerical features of the dataset were used. Since the dataset was already scaled to 20000 rows during the preprocessing phase, there was no need to scale it further.

### 6.1 Clustering Method & Resulting Dendrogram

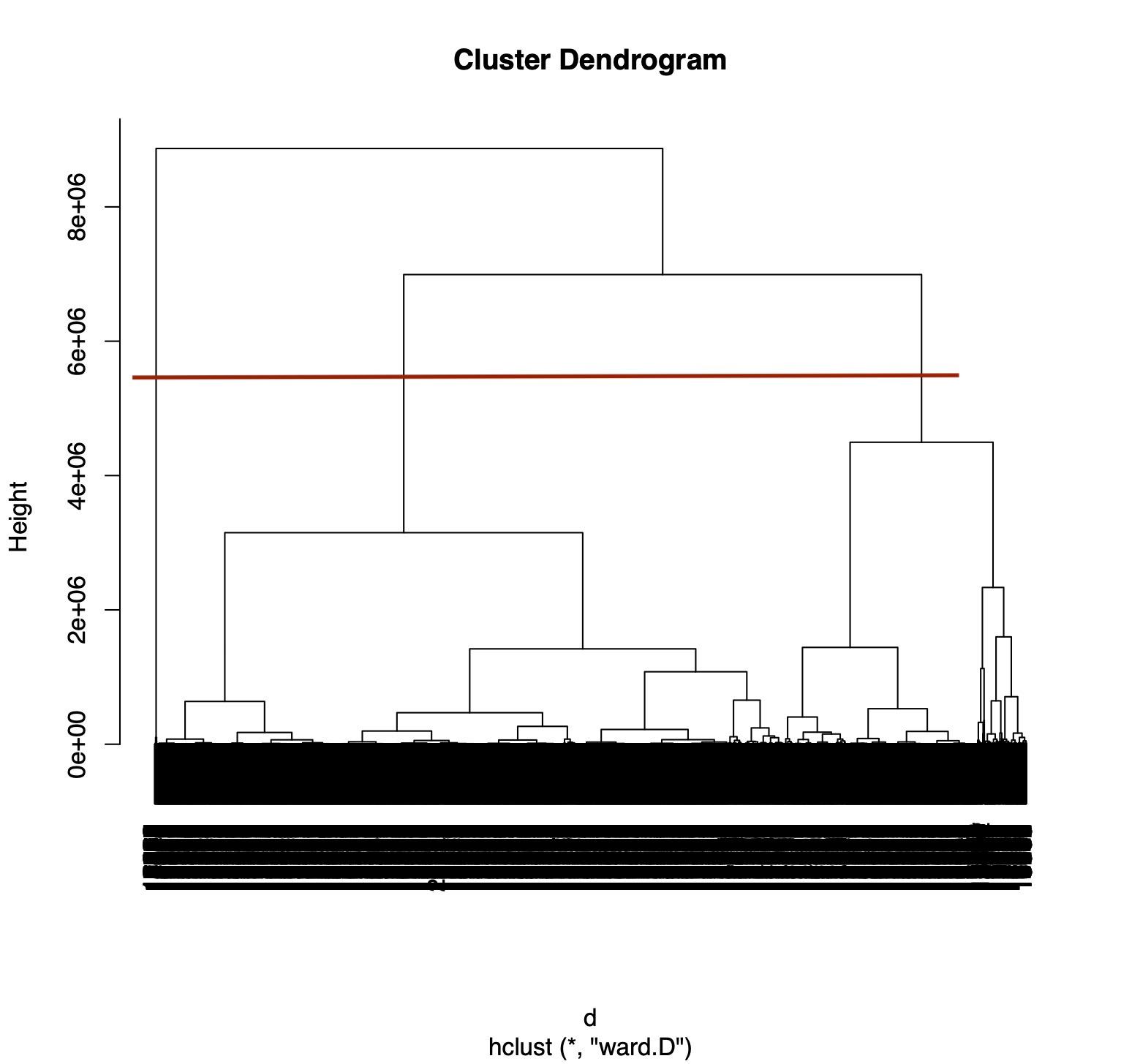
Two dendrograms were created with two different methods:

* method="ward.D"
* method="ward.D2" + dissimilarity matrix

" ward.D" = Ward’s minimum variance method

" ward.D2" = Ward’s minimum variance method – however dissimilarities are squared before clustering

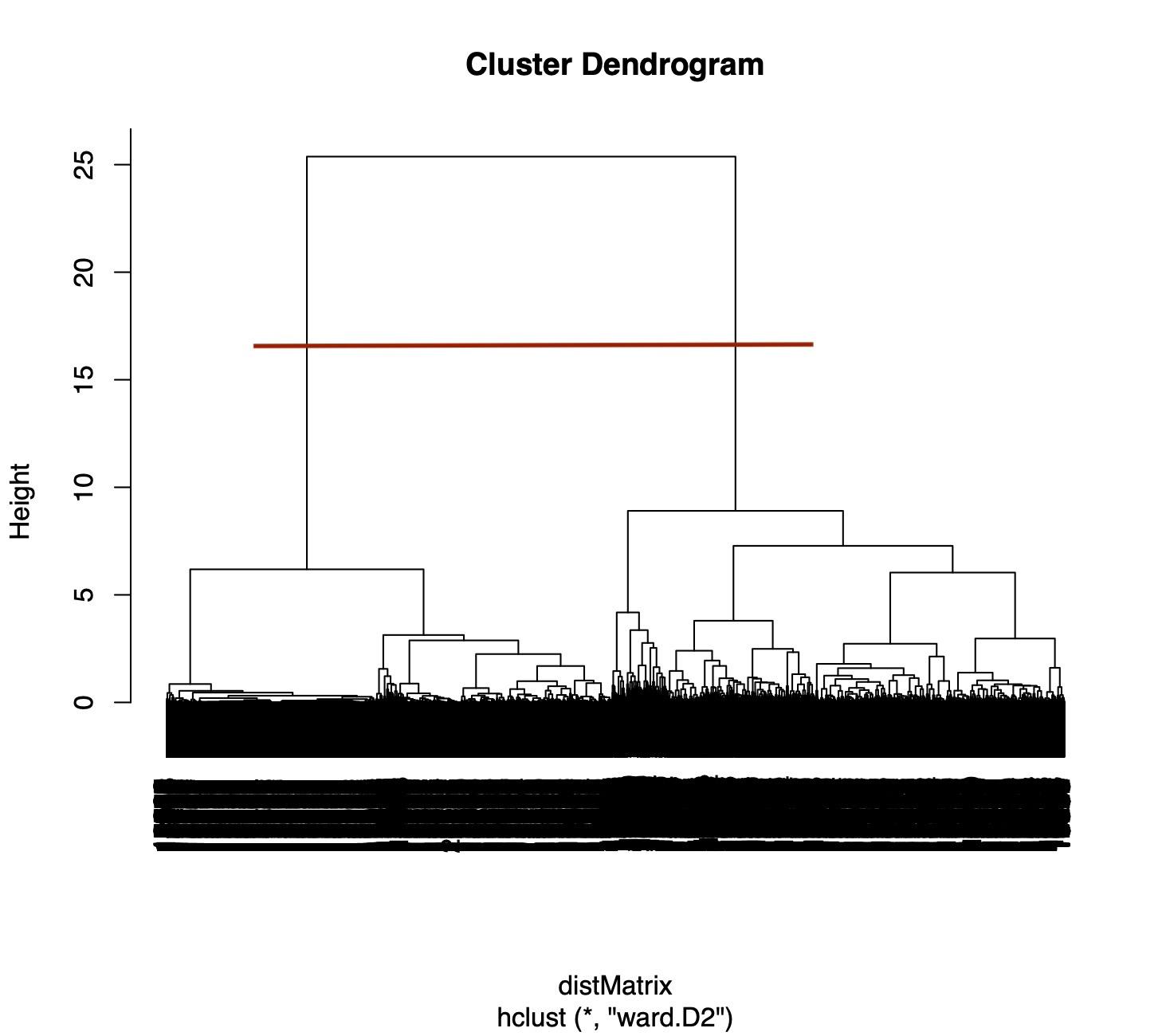
### 



### 6.1.1 Dendrogram method="ward.D"

Considering the distances between clusters deduced from the heights of the vertical lines of the graph, the result with 3 clusters is the most correct, but we will see later how this last one doesn't make much sense and how one of the clusters is very small and groups only marginal values

### 



### 6.1.2 Dendrogram method="ward.D2", dissimilarity matrix

In this dendrogram the distinction between clusters is clear, we can distinguish two very balanced groups. Considering the ambiguity of the previous result, I decide to continue the analysis taking into account this last method.

### 

### 

### 6.2 Clusters Size

method="ward.D"

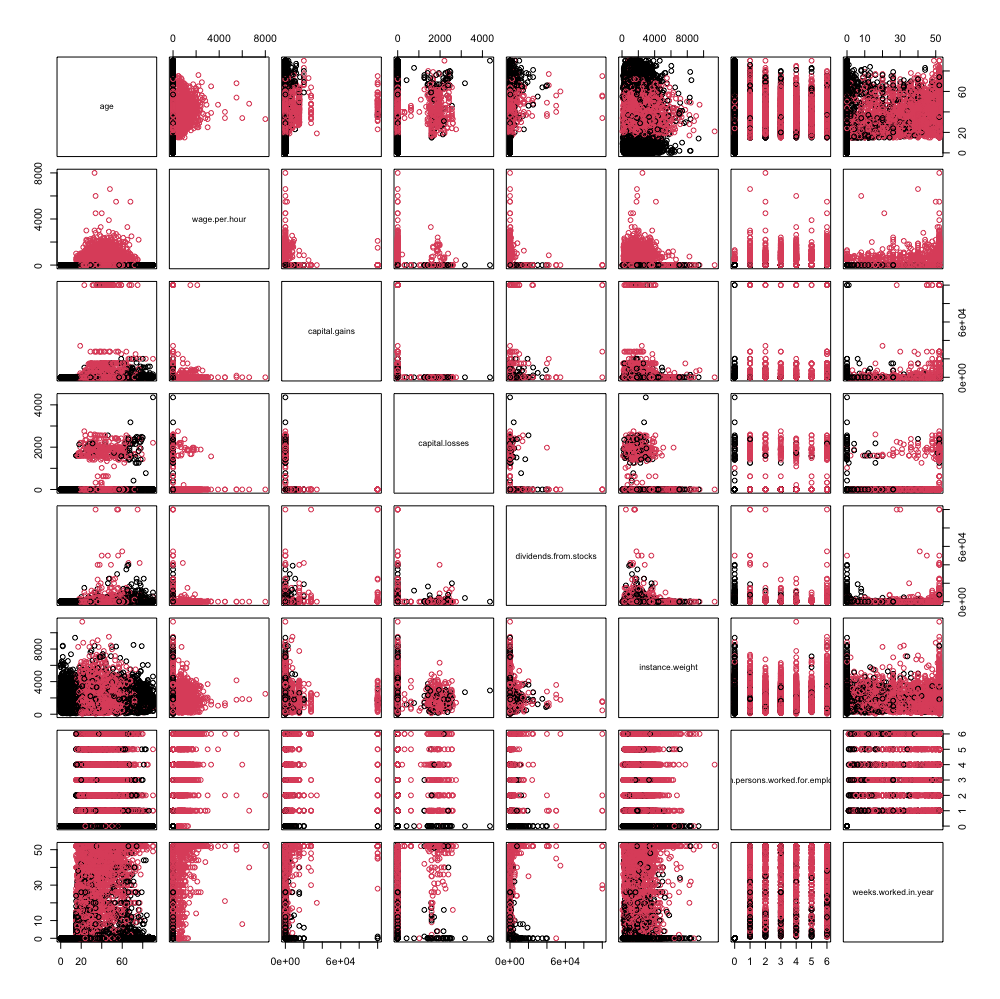
| 1 | 2 | 3 |
| --- | --- | --- |
| 14415 | 5539 | 46 |

method="ward.D2", dissimilarity matrix

| 1 | 2 |
| --- | --- |
| 9923 | 10077 |

As seen in the previous section, the division into three with "word.D" cluster is very unbalanced and probably a cluster takes into account only some marginal value that should be assigned to one of the other two clusters. On the contrary with "word.D2" the two clusters divide the dataset into almost equal parts without creating imbalances, for this reason it is preferable to continue taking into consideration the division into two clusters

### 6.3 Cluster analysis and Conclusion



Going to set the x as the axis that represents age we can immediately see two distinct clusters. The first ranging from 0 to 18 and from 65 onwards, and then the one in the middle, then from 18 to 65. This figure makes sense as it represents the distinction between workers and non-workers. It is also interesting to note the Gaussian trend that the wage per hour has in relation to the visible age of the last graph

