UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

FIB

Report

United States census income data

Data Mining 2022-2023 Q2

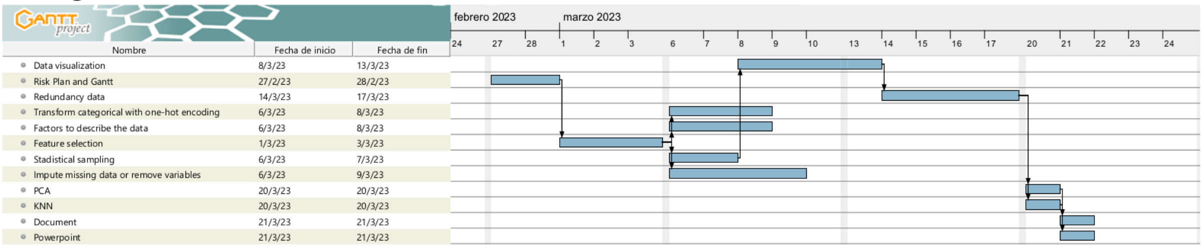
Authors:
Pol Casacuberta Gil
He Chen
Eric Hurtado
Tommaso Patriti
Alexandru-Ilie Popa

Facultat d'Informàtica de Barcelona

1.Initial working plan

Participant	Alex	Pol	Tommaso	Eric	He
1. Risk Plan and Gantt(2) D2		X			**X**
2. Preprocessing					
2.1 Feature selection (2) (D1)					**X**
2.2 Stadistical sampling (D2)					
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**	
3. PCA (4)	X	**X**			
4. KNN (5)			**X**		
5. Document (word style)	X	X	X	X	X
6. Powerpoint	X	X	X	X	X
DURATION: from shortest to longest D1-D5					

Workflow grid:



Nombre	Fecha de inicio	Fecha de fin
Data visualization	8/3/23	13/3/23
Risk Plan and Gantt	27/2/23	28/2/23
Redundancy data	14/3/23	17/3/23
Transform categorical with one-hot encoding	6/3/23	8/3/23
Factors to describe the data	6/3/23	8/3/23
Feature selection	1/3/23	3/3/23
Stadistical sampling	6/3/23	7/3/23
Impute missing data or remove variables	6/3/23	9/3/23
PCA	20/3/23	20/3/23
KNN	20/3/23	20/3/23
Document	21/3/23	21/3/23
Powerpoint	21/3/23	21/3/23

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	0	respondent’s industry

Variable	Type	Modalities	Missing percent	Meaning
		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		

Variable	Type	Modalities	Missing percent	Meaning
		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		
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	0	respondent’s occupation

Variable	Type	Modalities	Missing percent	Meaning
		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		

Variable	Type	Modalities	Missing percent	Meaning
		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		
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	0	respondent’s education

Variable	Type	Modalities	Missing percent	Meaning
		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		
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	0	respondent’s major industry

Variable	Type	Modalities	Missing percent	Meaning
		Manufacturing-nondurable goods Private household services Communications Mining Armed Forces Medical except hospital Entertainment Social services Utilities and sanitary services Forestry and fisheries		
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	0	respondent's race

Variable	Type	Modalities	Missing percent	Meaning
		Asian or Pacific Islander Other		
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	0	respondent’s tax filer stat

Variable	Type	Modalities	Missing percent	Meaning
		Single Head of household Joint both 65+ Joint one under 65 & one 65+		
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	0	respondent’s household and family stat

Variable	Type	Modalities	Missing percent	Meaning
		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		
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

Variable	Type	Modalities	Missing percent	Meaning
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	0	respondent’s country birth of father

Variable	Type	Modalities	Missing percent	Meaning
		Thailand South Korea Yugoslavia Hungary Vietnam Jamaica Columbia Holand-Netherlands Trinadad&Tobago Hong Kong Panama		
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	0	respondent’s country birth of mother

Variable	Type	Modalities	Missing percent	Meaning
		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		
country.of.birth.self	categorical	United-States Mexico Taiwan China Ireland Canada	0	respondent’s country birth

Variable	Type	Modalities	Missing percent	Meaning
		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		

Variable	Type	Modalities	Missing percent	Meaning
		Trinadad&Tobago Portugal Cambodia Holand-Netherlands Hungary		
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	A broad code indicating the major occupation in which the individual	Categorical	50.4623527

	code	is employed.		
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 /	These three features were analyzed together because they are closely related. If the father was born in country X, there is	New categories created:

	country of birth self	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.	"UnknownFather Country", "UnknownMother Country", "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	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	Remove

	income-related benefits)	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.	

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:

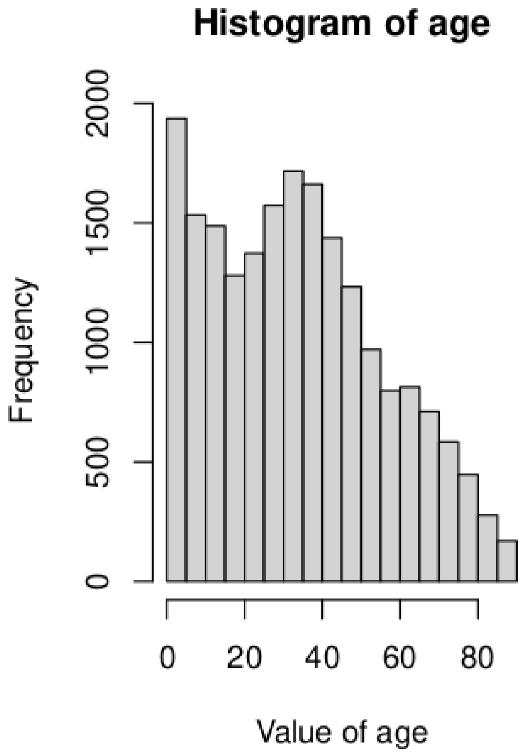
- For categorical variables, a pie chart and bar plot were computed. Only legible pie charts were included. Subsequently, the quantity of distinct modalities is displayed, followed by a tabulation containing the count for each modality and the relative frequency of each factor expressed as a proportion.
- 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

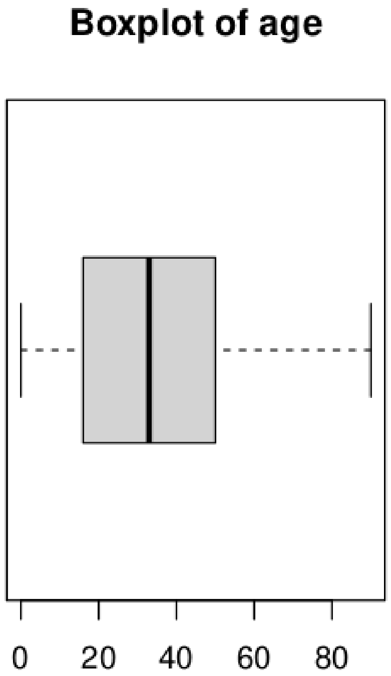
well as the mean, standard deviation, coefficient of variation and quantity of unknown values.

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

Variable 1 : age

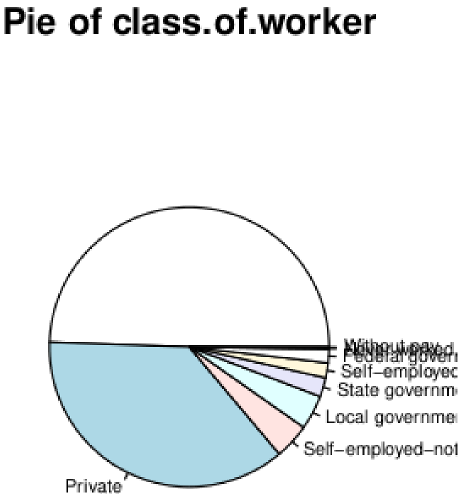


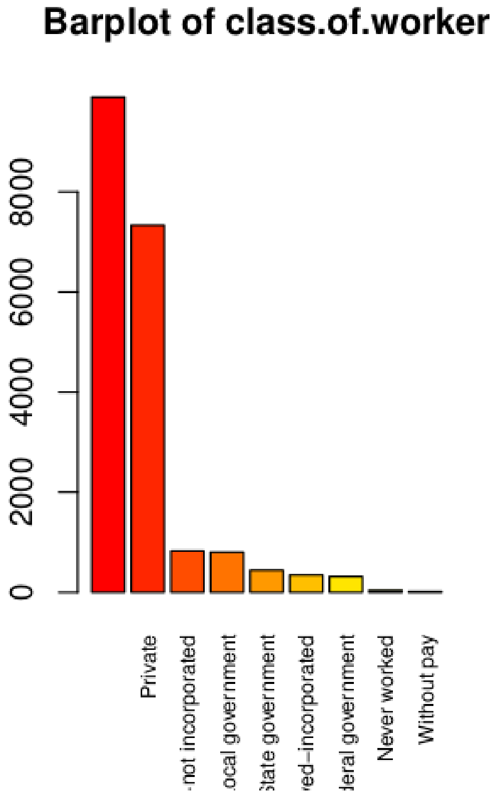


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	16	33	34.41035	50	90	22.11444	0.6426684	0

Table 6.3. age extended Summary Statistics.

Variable 2 : class.of.worker



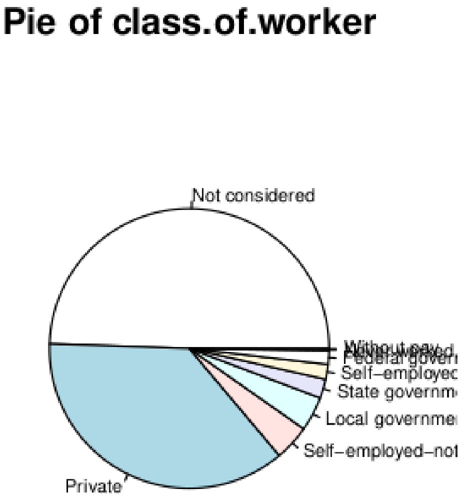


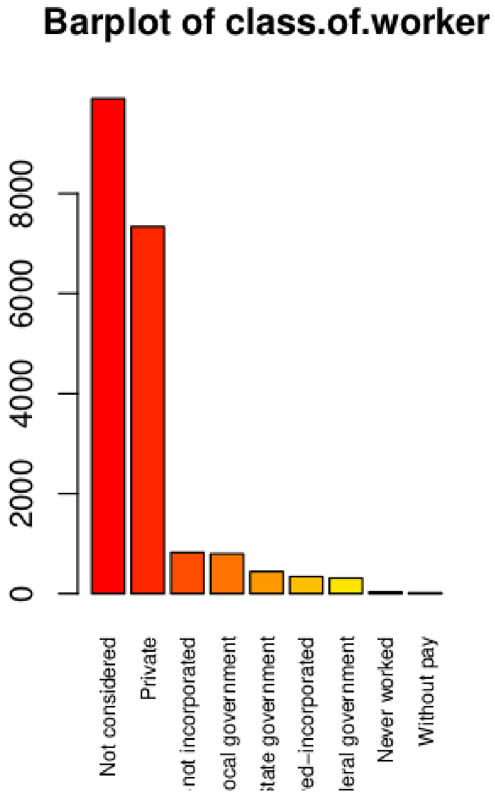
Number of modalities: 9

class.of.worker	Frequency	Proportion
NA	9892	0.49460
Private	7333	0.36665
Self-employed-not incorporated	826	0.04130
Local government	798	0.03990
State government	442	0.02210
Self-employed-incorporated	345	0.01725
Federal government	313	0.01565
Never worked	36	0.00180
Without pay	15	0.00075

Table 6.4. class.of.worker frequency and proportion table.

Variable 2 : class.of.worker (CHANGED in preprocessing)


Pie of class.of.worker

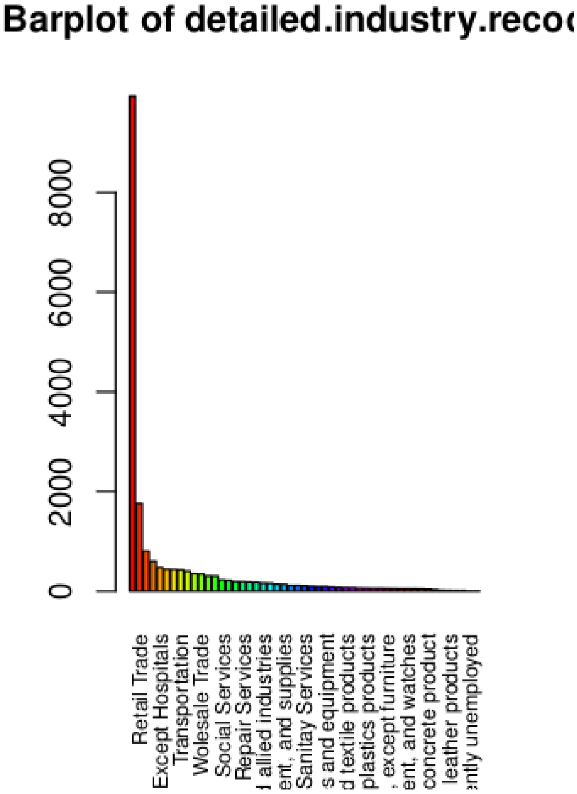

Barplot of class.of.worker

Number of modalities: 9

class.of.worker	Frequency	Proportion
Not considered	9892	0.49460
Private	7333	0.36665
Self-employed-not incorporated	826	0.04130
Local government	798	0.03990
State government	442	0.02210
Self-employed-incorporated	345	0.01725
Federal government	313	0.01565
Never worked	36	0.00180
Without pay	15	0.00075

Table 6.5. class.of.worker frequency and proportion table.

Variable 3 : detailed.industry.recode

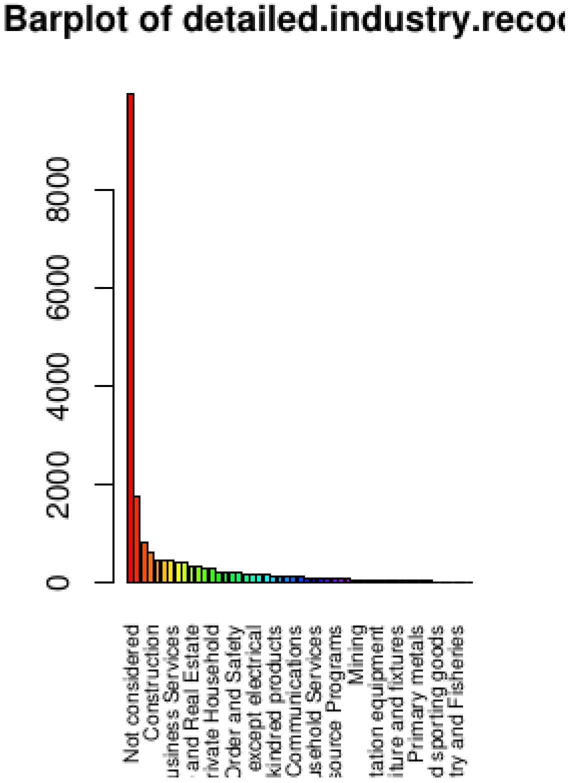


Number of modalities: 51

detailed.industry.recode	Frequency	Proportion
NA	9928	0.49640
Retail Trade	1762	0.08810
Educational Services	803	0.04015
Construction	602	0.03010
Health Services, Except Hospitals	475	0.02375
Other Professional Services	441	0.02205
Business Services	437	0.02185
Transportation	429	0.02145
Hospitals	401	0.02005
Insurance and Real Estate	352	0.01760
Wolesale Trade	345	0.01725
Banking and Other Finance	304	0.01520
Personal Services, Except Private Household	301	0.01505
Social Services	229	0.01145
Other Agriculture	219	0.01095
Justice, Public Order and Safety	194	0.00970
Repair Services	192	0.00960
Other Public Administration	185	0.00925
Machinery, except electrical	181	0.00905
Printing, publishing and allied industries	166	0.00830
Entertainment and Recreation Services	162	0.00810
Food and kindred products	142	0.00710
Electrical machinery, equipment, and supplies	138	0.00690
Chemicals and allied products	116	0.00580
Communications	116	0.00580
Utilities and Sanitay Services	111	0.00555
Fabricated metal	99	0.00495
Private Household Services	95	0.00475
Motor vehicles and equipment	94	0.00470
Agriculture Service	84	0.00420
Administration of Human Resource Programs	74	0.00370
Apparel and other finished textile products	70	0.00350
National Security and Internal Affairs	69	0.00345
Mining	62	0.00310
Rubber and miscellaneous plastics products	58	0.00290
Paper and allied products	57	0.00285
Other transportation equipment	56	0.00280
Lumber and wood products, except furniture	54	0.00270
Miscellaneous and not specified manufacturing industries	53	0.00265
Furniture and fixtures	51	0.00255
Professional and photographic equipment, and watches	51	0.00255
Textile mill products	49	0.00245
Primary metals	48	0.00240
Stone clay, glass, and concrete product	44	0.00220
Aircraft and parts	32	0.00160
Toys, amusements, and sporting goods	18	0.00090
Leather and leather products	15	0.00075
Petroleum and coal products	14	0.00070
Forestry and Fisheries	13	0.00065
Armed Forces last job, currently unemployed	5	0.00025
Tobacco manufactures	4	0.00020

Table 6.6. detailed.industry.recode frequency and proportion table.

Variable 3 : detailed.industry.recode (CHANGED in preprocessing)

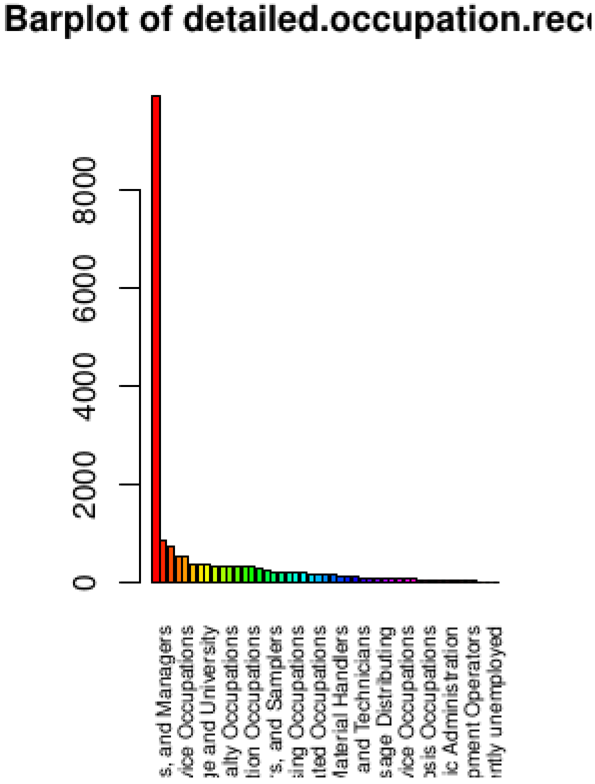


Number of modalities: 51

detailed.industry.recode	Frequency	Proportion
Not considered	9928	0.49640
Retail Trade	1762	0.08810
Educational Services	803	0.04015
Construction	602	0.03010
Health Services, Except Hospitals	475	0.02375
Other Professional Services	441	0.02205
Business Services	437	0.02185
Transportation	429	0.02145
Hospitals	401	0.02005
Insurance and Real Estate	352	0.01760
Wolesale Trade	345	0.01725
Banking and Other Finance	304	0.01520
Personal Services, Except Private Household	301	0.01505
Social Services	229	0.01145
Other Agriculture	219	0.01095
Justice, Public Order and Safety	194	0.00970
Repair Services	192	0.00960
Other Public Administration	185	0.00925
Machinery, except electrical	181	0.00905
Printing, publishing and allied industries	166	0.00830
Entertainment and Recreation Services	162	0.00810
Food and kindred products	142	0.00710
Electrical machinery, equipment, and supplies	138	0.00690
Chemicals and allied products	116	0.00580
Communications	116	0.00580
Utilities and Sanitay Services	111	0.00555
Fabricated metal	99	0.00495
Private Household Services	95	0.00475
Motor vehicles and equipment	94	0.00470
Agriculture Service	84	0.00420
Administration of Human Resource Programs	74	0.00370
Apparel and other finished textile products	70	0.00350
National Security and Internal Affairs	69	0.00345
Mining	62	0.00310
Rubber and miscellaneous plastics products	58	0.00290
Paper and allied products	57	0.00285
Other transportation equipment	56	0.00280
Lumber and wood products, except furniture	54	0.00270
Miscellaneous and not specified manufacturing industries	53	0.00265
Furniture and fixtures	51	0.00255
Professional and photographic equipment, and watches	51	0.00255
Textile mill products	49	0.00245
Primary metals	48	0.00240
Stone clay, glass, and concrete product	44	0.00220
Aircraft and parts	32	0.00160
Toys, amusements, and sporting goods	18	0.00090
Leather and leather products	15	0.00075
Petroleum and coal products	14	0.00070
Forestry and Fisheries	13	0.00065
Armed Forces last job, currently unemployed	5	0.00025
Tobacco manufactures	4	0.00020

Table 6.7. detailed.industry.recode frequency and proportion table.

Variable 4 : detailed.occupation.recode

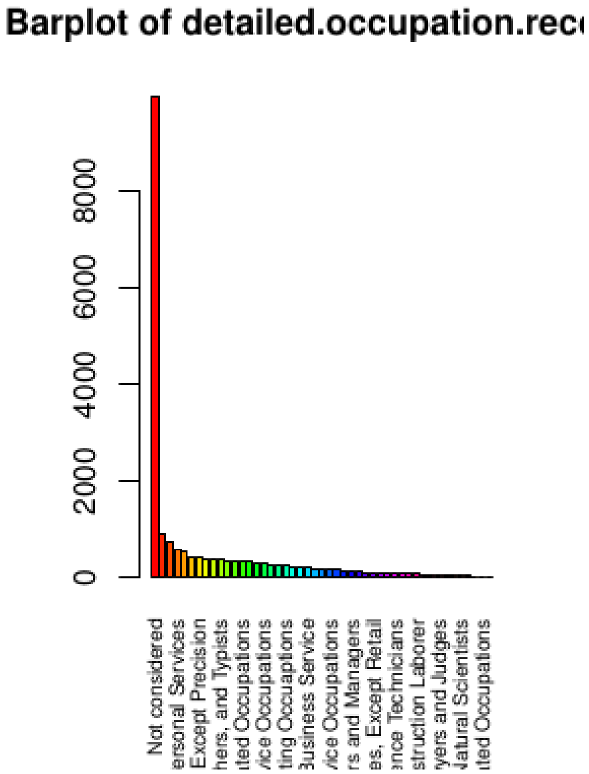


Number of modalities: 47

detailed.occupation.recode	Frequency	Proportion
NA	9928	0.49640
Other Executive, Administrators, and Managers	894	0.04470
Other Administrative Support Occupations, Including Clerical	758	0.03790
Sales Workers, Retail and Personal Services	565	0.02825
Food Service Occupations	549	0.02745
Construction Trades	406	0.02030
Machine Operators and Tenders, Except Precision	403	0.02015
Teachers, Except College and University	368	0.01840
Mechanics and Repairers	357	0.01785
Secretaries, Stenographers, and Typists	356	0.01780
Other Professional Specialty Occupations	351	0.01755
Supervisors and Proprietors, Sales Occupations	341	0.01705
Management Related Occupations	329	0.01645
Other Precision Production Occupations	329	0.01645
Motor Vehicle Operators	290	0.01450
Cleaning and Building Service Occupations	278	0.01390
Fabricators, Assemblers, Inspectors, and Samplers	240	0.01200
Personal Service Occupations	237	0.01185
Health Assessment and Treating Occuaptions	235	0.01175
Financial Records, Processing Occupations	206	0.01030
Other Handlers, Equipment Cleaners, and Laborers	205	0.01025
Sales Representatives, Finance, and Business Service	193	0.00965
Farm Workers and Related Occupations	179	0.00895
Protective Service Occupations	173	0.00865
Health Service Occupations	167	0.00835
Freight, Stock and Material Handlers	158	0.00790
Engineers	148	0.00740
Farm Operators and Managers	123	0.00615
Health Technologists and Technicians	112	0.00560
Other Transportation Occupations and Material Moving	104	0.00520
Sales Representatives, Commodities, Except Retail	101	0.00505
Mail and Message Distributing	92	0.00460
Technicians, Except Health, Engineering, and Science	85	0.00425
Engineering and Science Technicians	84	0.00420
Private Household Service Occupations	83	0.00415
Mathematical and Computer Scientists	82	0.00410
Construction Laborer	70	0.00350
Health Diagnosis Occupations	68	0.00340
Teachers, College and University	67	0.00335
Lawyers and Judges	58	0.00290
Public Administration	55	0.00275
Supervisors - Administrative Support	53	0.00265
Natural Scientists	49	0.00245
Computer Equipment Operators	43	0.00215
Forestry and Fishing Occupations	16	0.00080
Sales Related Occupations	7	0.00035
Armed Forces last job, currently unemployed	5	0.00025

Table 6.8. detailed.occupation.recode frequency and proportion table.

Variable 4 : detailed.occupation.recode (CHANGED in preprocessing)

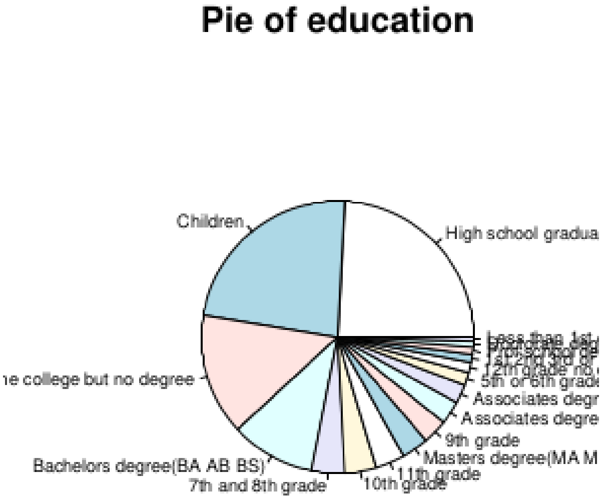


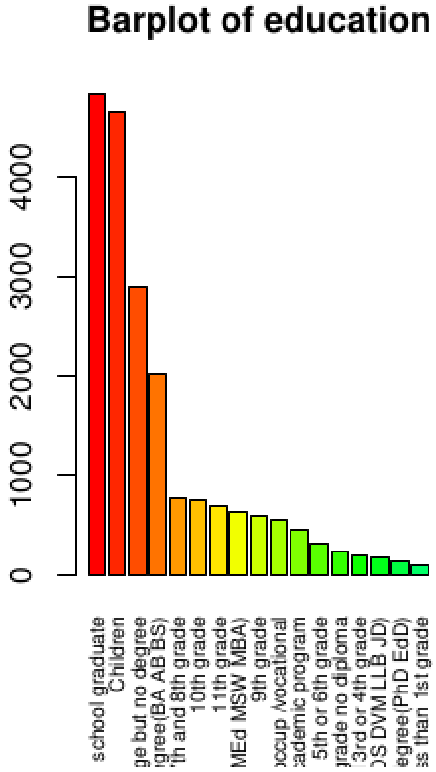
Number of modalities: 47

detailed.occupation.recode	Frequency	Proportion
Not considered	9928	0.49640
Other Executive, Administrators, and Managers	894	0.04470
Other Administrative Support Occupations, Including Clerical	758	0.03790
Sales Workers, Retail and Personal Services	565	0.02825
Food Service Occupations	549	0.02745
Construction Trades	406	0.02030
Machine Operators and Tenders, Except Precision	403	0.02015
Teachers, Except College and University	368	0.01840
Mechanics and Repairers	357	0.01785
Secretaries, Stenographers, and Typists	356	0.01780
Other Professional Specialty Occupations	351	0.01755
Supervisors and Proprietors, Sales Occupations	341	0.01705
Management Related Occupations	329	0.01645
Other Precision Production Occupations	329	0.01645
Motor Vehicle Operators	290	0.01450
Cleaning and Building Service Occupations	278	0.01390
Fabricators, Assemblers, Inspectors, and Samplers	240	0.01200
Personal Service Occupations	237	0.01185
Health Assessment and Treating Occuaptions	235	0.01175
Financial Records, Processing Occupations	206	0.01030
Other Handlers, Equipment Cleaners, and Laborers	205	0.01025
Sales Representatives, Finance, and Business Service	193	0.00965
Farm Workers and Related Occupations	179	0.00895
Protective Service Occupations	173	0.00865
Health Service Occupations	167	0.00835
Freight, Stock and Material Handlers	158	0.00790
Engineers	148	0.00740
Farm Operators and Managers	123	0.00615
Health Technologists and Technicians	112	0.00560
Other Transportation Occupations and Material Moving	104	0.00520
Sales Representatives, Commodities, Except Retail	101	0.00505
Mail and Message Distributing	92	0.00460
Technicians, Except Health, Engineering, and Science	85	0.00425
Engineering and Science Technicians	84	0.00420
Private Household Service Occupations	83	0.00415
Mathematical and Computer Scientists	82	0.00410
Construction Laborer	70	0.00350
Health Diagnosis Occupations	68	0.00340
Teachers, College and University	67	0.00335
Lawyers and Judges	58	0.00290
Public Administration	55	0.00275
Supervisors - Administrative Support	53	0.00265
Natural Scientists	49	0.00245
Computer Equipment Operators	43	0.00215
Forestry and Fishing Occupations	16	0.00080
Sales Related Occupations	7	0.00035
Armed Forces last job, currently unemployed	5	0.00025

Table 6.9. detailed.occupation.recode frequency and proportion table.

Variable 5 : education



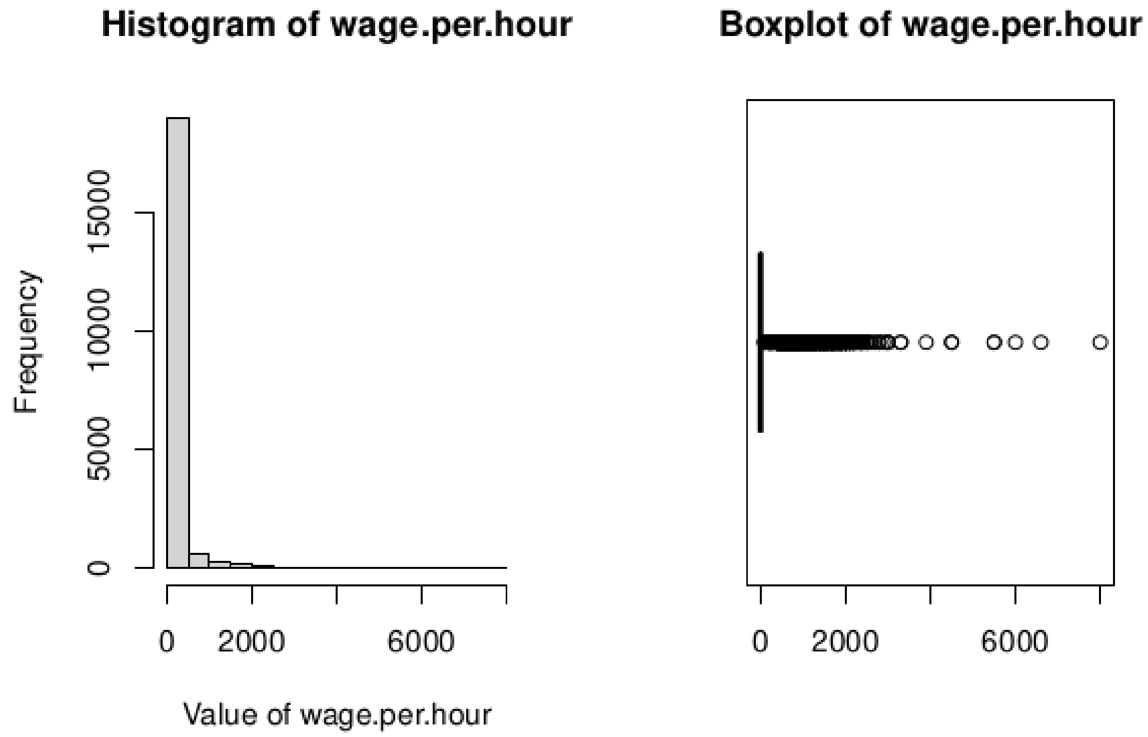


Number of modalities: 17

education	Frequency	Proportion
High school graduate	4832	0.24160
Children	4652	0.23260
Some college but no degree	2892	0.14460
Bachelors degree(BA AB BS)	2011	0.10055
7th and 8th grade	781	0.03905
10th grade	749	0.03745
11th grade	701	0.03505
Masters degree(MA MS MEng MEd MSW MBA)	640	0.03200
9th grade	592	0.02960
Associates degree-occup /vocational	555	0.02775
Associates degree-academic program	448	0.02240
5th or 6th grade	308	0.01540
12th grade no diploma	243	0.01215
1st 2nd 3rd or 4th grade	189	0.00945
Prof school degree (MD DDS DVM LLB JD)	169	0.00845
Doctorate degree(PhD EdD)	147	0.00735
Less than 1st grade	91	0.00455

Table 6.10. education frequency and proportion table.

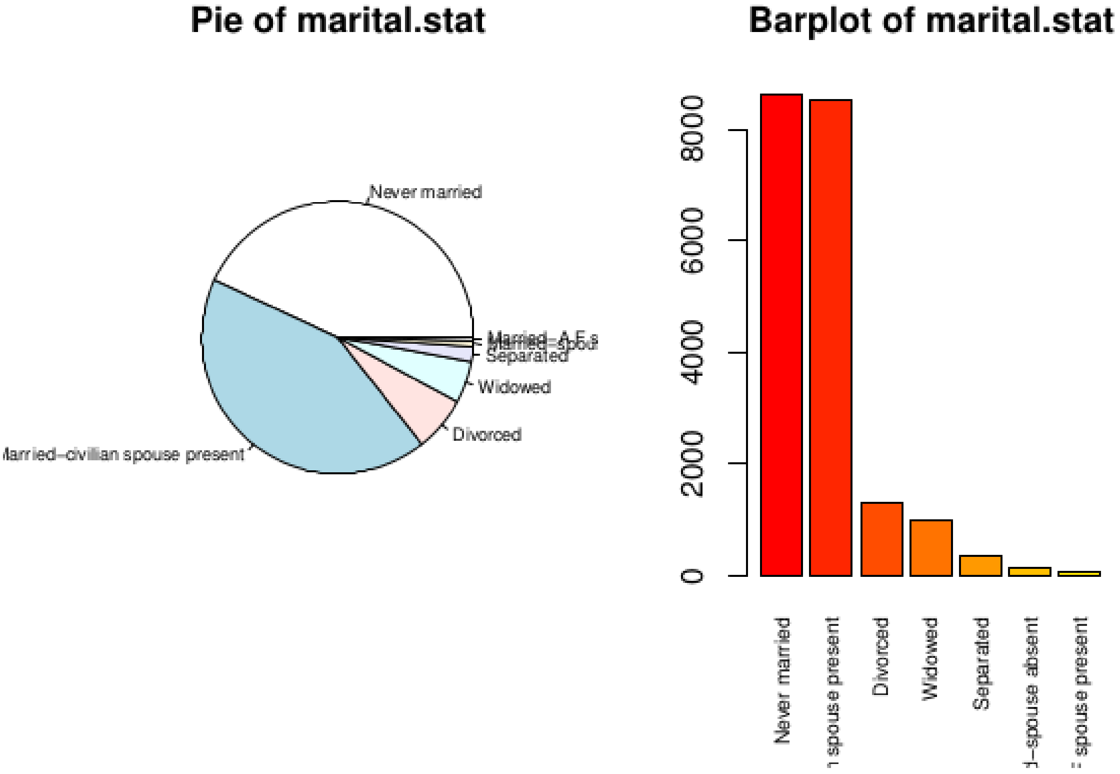
Variable 6 : wage.per.hour



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	59.62675	0	8000	282.2555	4.733706	0

Table 6.11. wage.per.hour extended Summary Statistics.

Variable 7 : marital.stat

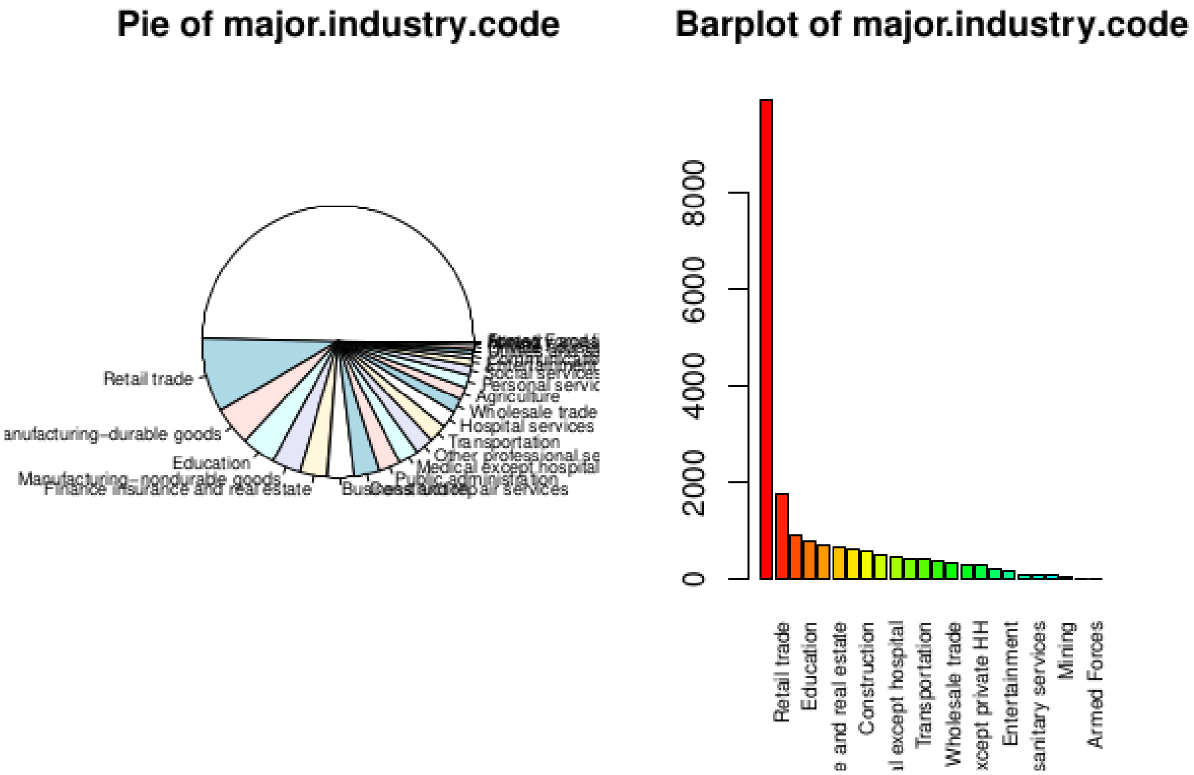


Number of modalities: 7

marital.stat	Frequency	Proportion
Never married	8618	0.4309
Married-civilian spouse present	8514	0.4257
Divorced	1302	0.0651
Widowed	992	0.0496
Separated	350	0.0175
Married-spouse absent	148	0.0074
Married-A F spouse present	76	0.0038

Table 6.12. marital.stat frequency and proportion table.

Variable 8 : major.industry.code

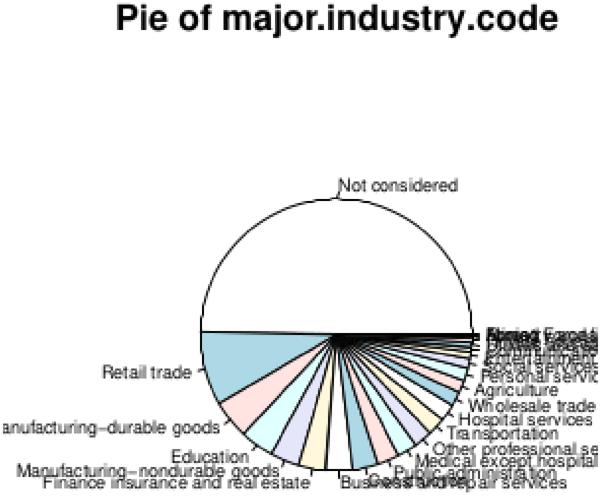


Number of modalities: 24

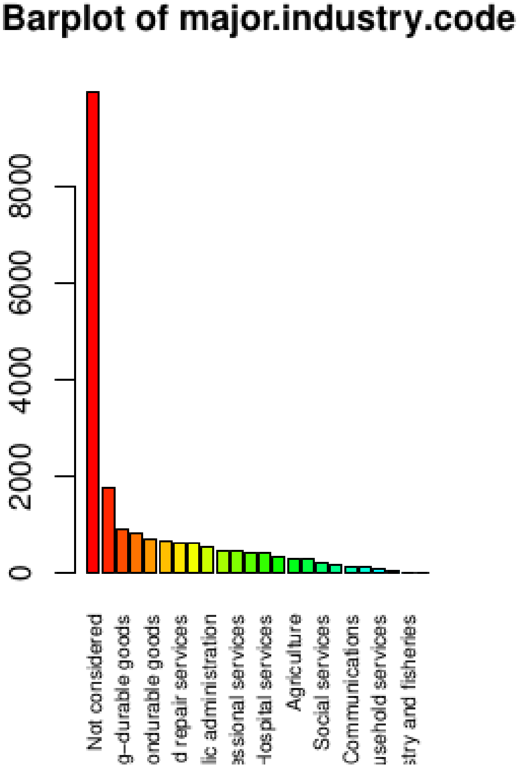
major.industry.code	Frequency	Proportion
NA	9928	0.49640
Retail trade	1762	0.08810
Manufacturing-durable goods	919	0.04595
Education	803	0.04015
Manufacturing-nondurable goods	691	0.03455
Finance insurance and real estate	656	0.03280
Business and repair services	629	0.03145
Construction	602	0.03010
Public administration	522	0.02610
Medical except hospital	475	0.02375
Other professional services	441	0.02205
Transportation	429	0.02145
Hospital services	401	0.02005
Wholesale trade	345	0.01725
Agriculture	303	0.01515
Personal services except private HH	301	0.01505
Social services	229	0.01145
Entertainment	162	0.00810
Communications	116	0.00580
Utilities and sanitary services	111	0.00555
Private household services	95	0.00475
Mining	62	0.00310
Forestry and fisheries	13	0.00065
Armed Forces	5	0.00025

Table 6.13. major.industry.code frequency and proportion table.

Variable 8 : major.industry.code (CHANGED in preprocessing)



Pie of major.industry.code



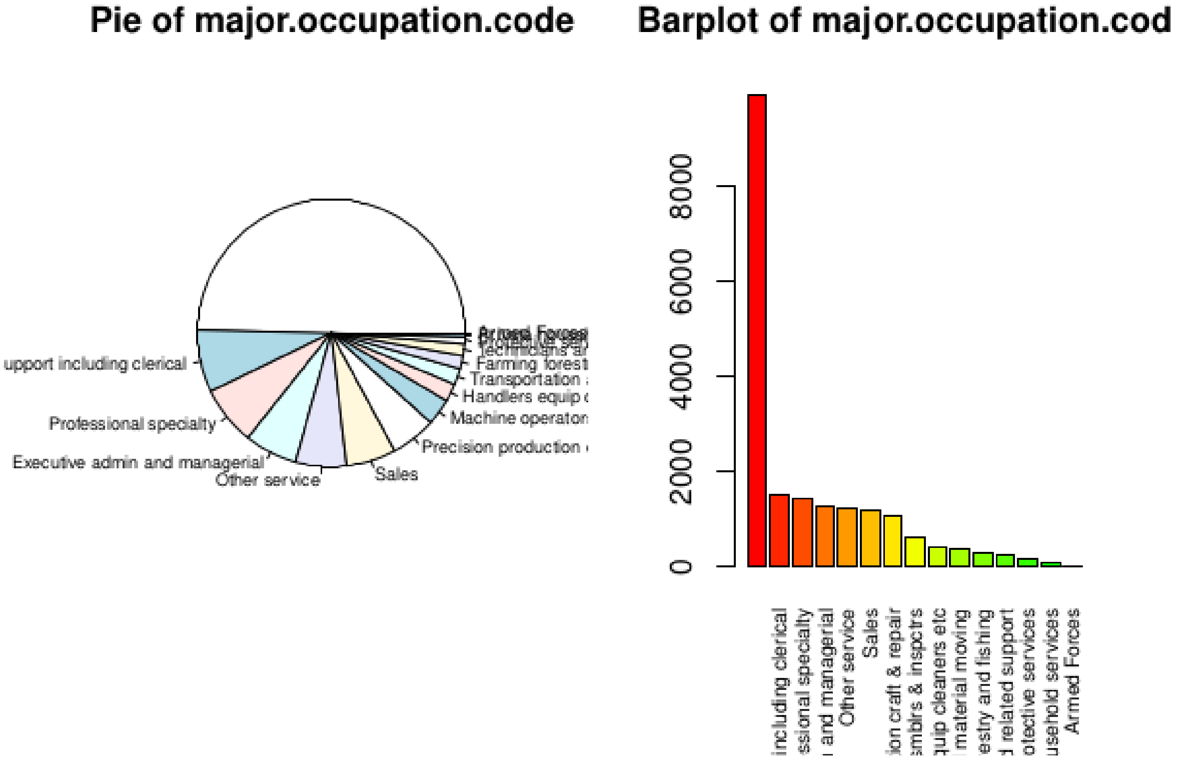
Barplot of major.industry.code

Number of modalities: 24

major.industry.code	Frequency	Proportion
Not considered	9928	0.49640
Retail trade	1762	0.08810
Manufacturing-durable goods	919	0.04595
Education	803	0.04015
Manufacturing-nondurable goods	691	0.03455
Finance insurance and real estate	656	0.03280
Business and repair services	629	0.03145
Construction	602	0.03010
Public administration	522	0.02610
Medical except hospital	475	0.02375
Other professional services	441	0.02205
Transportation	429	0.02145
Hospital services	401	0.02005
Wholesale trade	345	0.01725
Agriculture	303	0.01515
Personal services except private HH	301	0.01505
Social services	229	0.01145
Entertainment	162	0.00810
Communications	116	0.00580
Utilities and sanitary services	111	0.00555
Private household services	95	0.00475
Mining	62	0.00310
Forestry and fisheries	13	0.00065
Armed Forces	5	0.00025

Table 6.14. major.industry.code frequency and proportion table.

Variable 9 : major.occupation.code

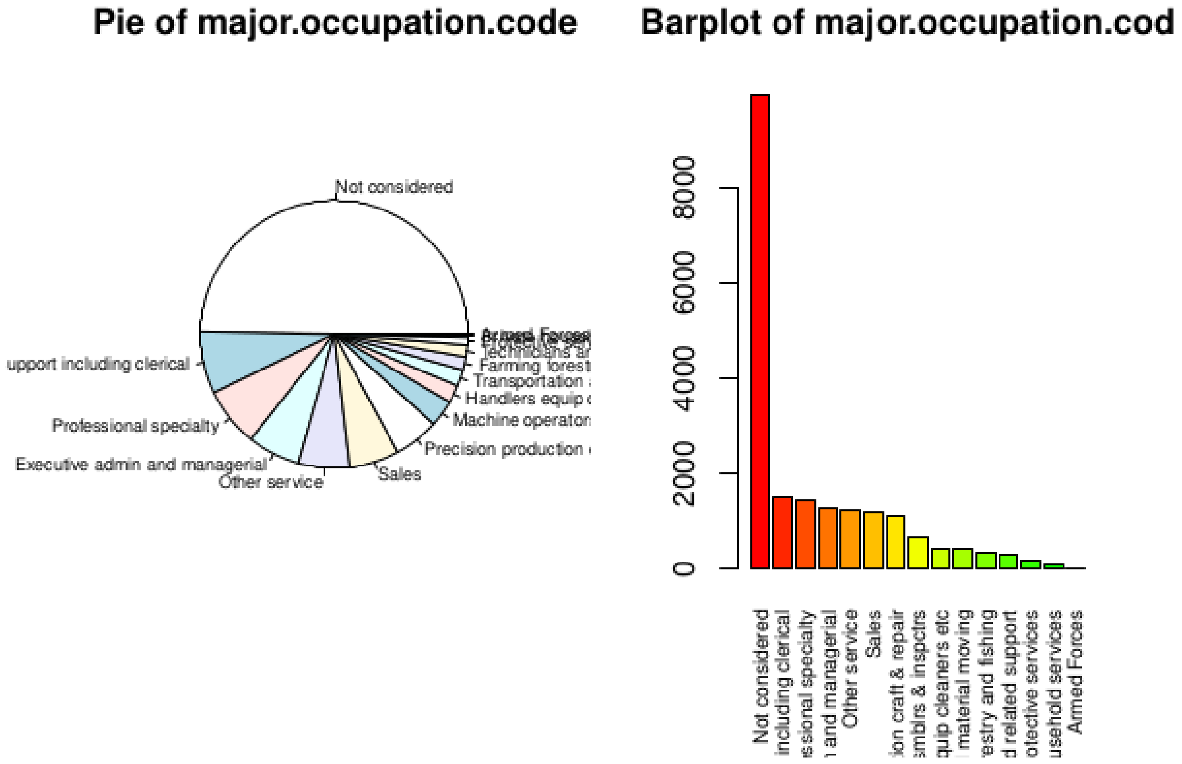


Number of modalities: 15

major.occupation.code	Frequency	Proportion
NA	9928	0.49640
Adm support including clerical	1508	0.07540
Professional specialty	1426	0.07130
Executive admin and managerial	1278	0.06390
Other service	1231	0.06155
Sales	1207	0.06035
Precision production craft & repair	1092	0.05460
Machine operators assmblrs & inspctrs	643	0.03215
Handlers equip cleaners etc	433	0.02165
Transportation and material moving	394	0.01970
Farming forestry and fishing	318	0.01590
Technicians and related support	281	0.01405
Protective services	173	0.00865
Private household services	83	0.00415
Armed Forces	5	0.00025

Table 6.15. major.occupation.code frequency and proportion table.

Variable 9 : major.occupation.code (CHANGED in preprocessing)


Pie of major.occupation.code

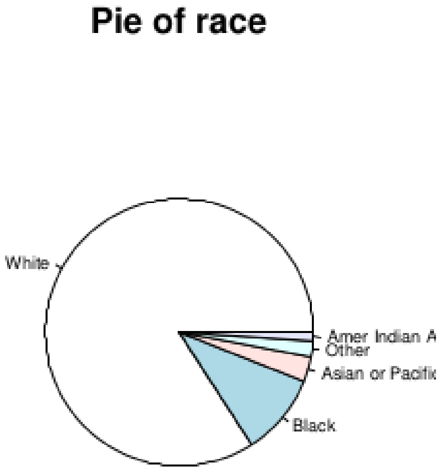
Barplot of major.occupation.cod

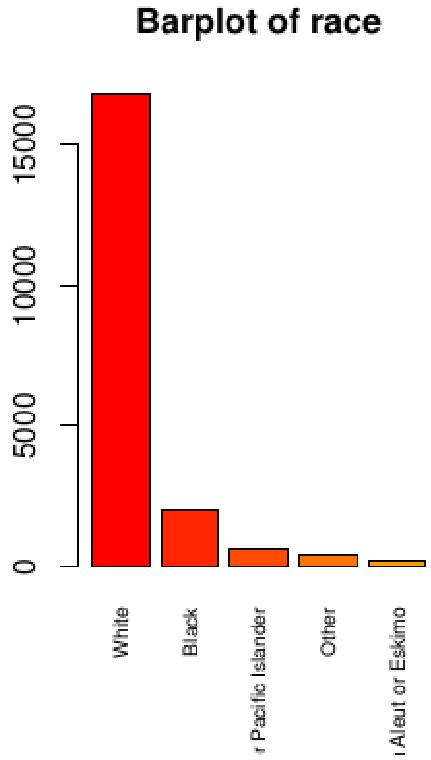
Number of modalities: 15

major.occupation.code	Frequency	Proportion
Not considered	9928	0.49640
Adm support including clerical	1508	0.07540
Professional specialty	1426	0.07130
Executive admin and managerial	1278	0.06390
Other service	1231	0.06155
Sales	1207	0.06035
Precision production craft & repair	1092	0.05460
Machine operators assmblrs & inspctrs	643	0.03215
Handlers equip cleaners etc	433	0.02165
Transportation and material moving	394	0.01970
Farming forestry and fishing	318	0.01590
Technicians and related support	281	0.01405
Protective services	173	0.00865
Private household services	83	0.00415
Armed Forces	5	0.00025

Table 6.16. major.occupation.code frequency and proportion table.

Variable 10 : race



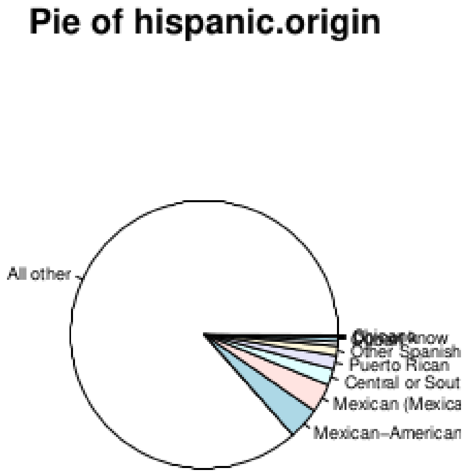


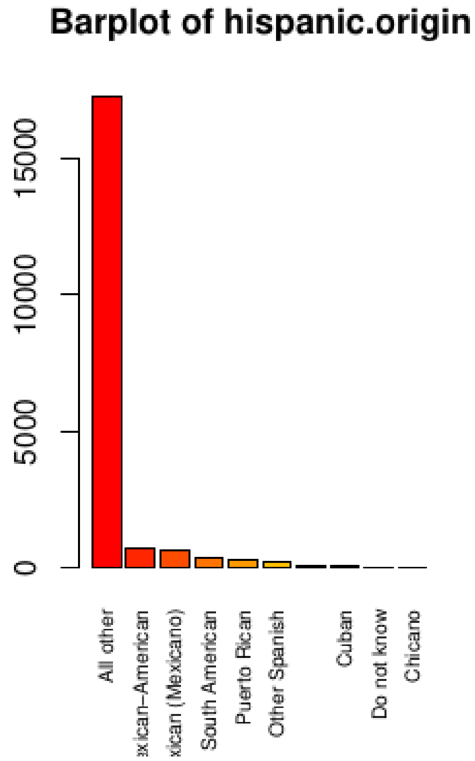
Number of modalities: 5

race	Frequency	Proportion
White	16818	0.84090
Black	1988	0.09940
Asian or Pacific Islander	604	0.03020
Other	381	0.01905
Amer Indian Aleut or Eskimo	209	0.01045

Table 6.17. race frequency and proportion table.

Variable 11 : hispanic.origin



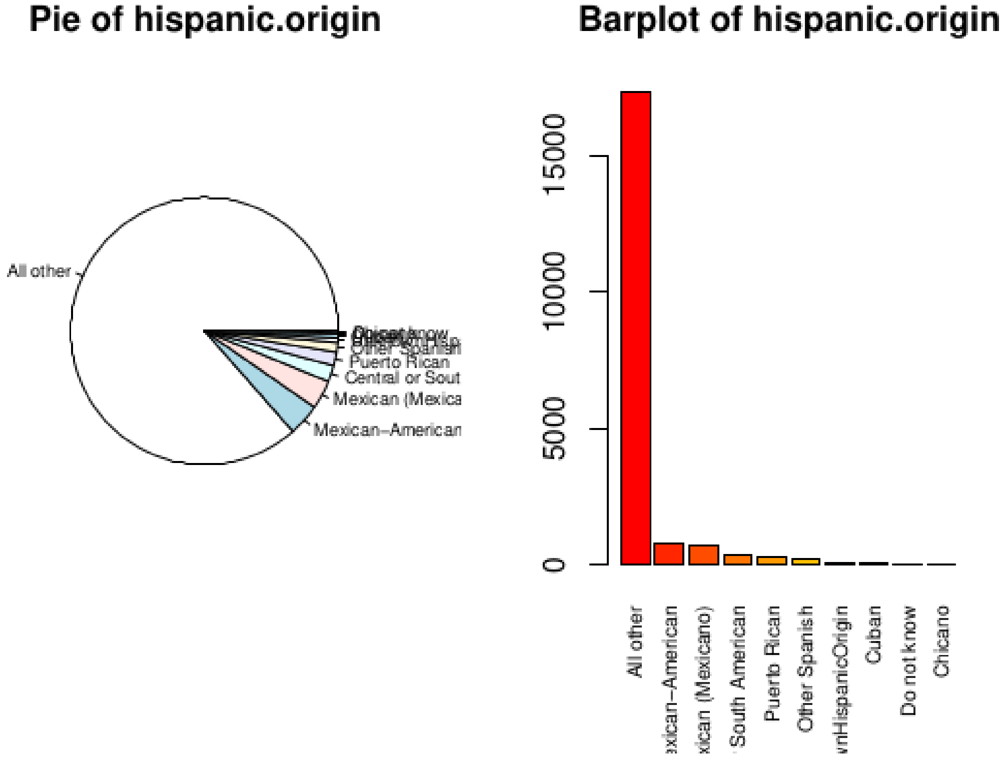


Number of modalities: 10

hispanic.origin	Frequency	Proportion
All other	17296	0.86480
Mexican-American	773	0.03865
Mexican (Mexicano)	697	0.03485
Central or South American	398	0.01990
Puerto Rican	327	0.01635
Other Spanish	236	0.01180
NA	103	0.00515
Cuban	101	0.00505
Do not know	35	0.00175
Chicano	34	0.00170

Table 6.18. hispanic.origin frequency and proportion table.

Variable 11 : hispanic.origin (CHANGED in preprocessing)

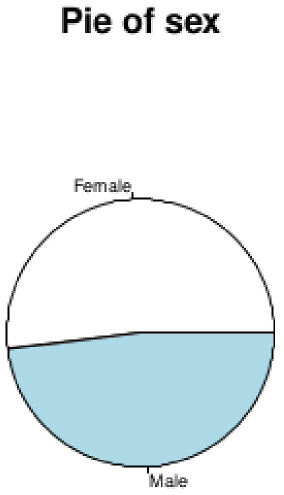


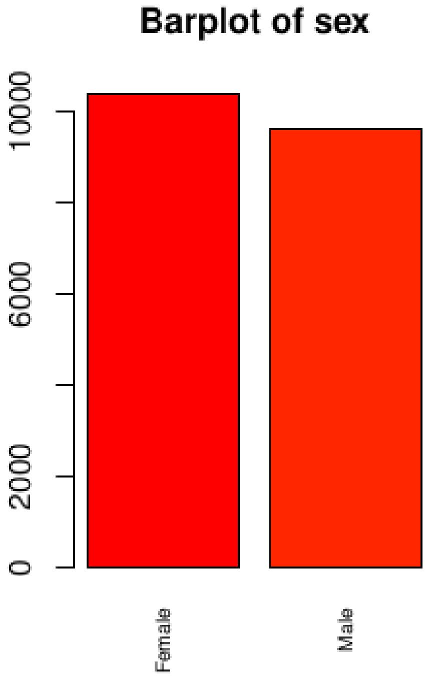
Number of modalities: 10

hispanic.origin	Frequency	Proportion
All other	17296	0.86480
Mexican-American	773	0.03865
Mexican (Mexicano)	697	0.03485
Central or South American	398	0.01990
Puerto Rican	327	0.01635
Other Spanish	236	0.01180
UnknownHispanicOrigin	103	0.00515
Cuban	101	0.00505
Do not know	35	0.00175
Chicano	34	0.00170

Table 6.19. hispanic.origin frequency and proportion table.

Variable 12 : sex



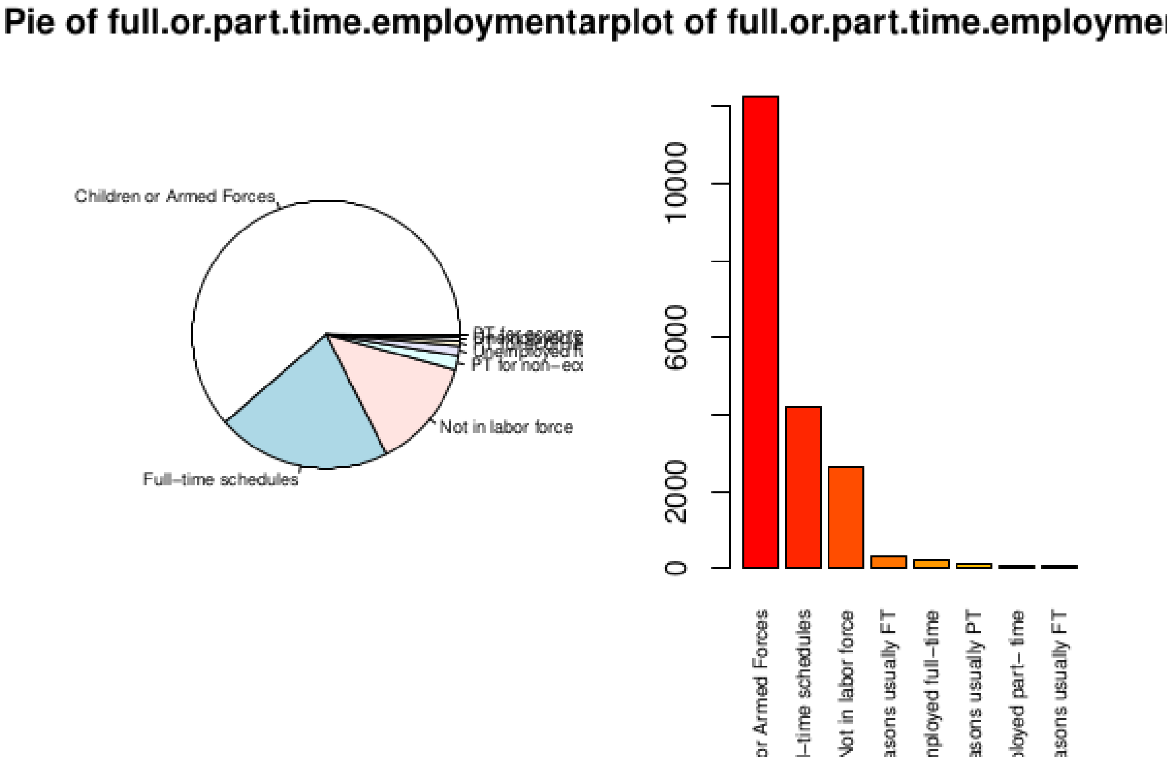


Number of modalities: 2

sex	Frequency	Proportion
Female	10383	0.51915
Male	9617	0.48085

Table 6.20. sex frequency and proportion table.

Variable 13 : full.or.part.time.employment.stat

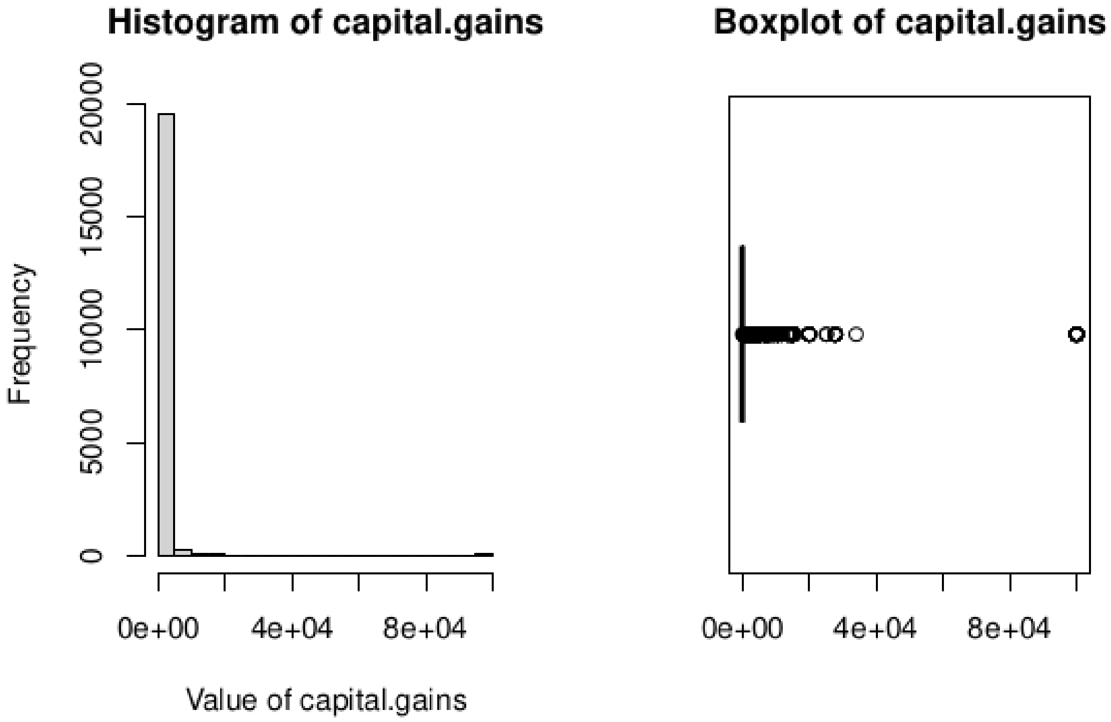


Number of modalities: 8

full.or.part.time.employment.stat	Frequency	Proportion
Children or Armed Forces	12278	0.6139
Full-time schedules	4212	0.2106
Not in labor force	2660	0.1330
PT for non-econ reasons usually FT	338	0.0169
Unemployed full-time	242	0.0121
PT for econ reasons usually PT	116	0.0058
Unemployed part- time	96	0.0048
PT for econ reasons usually FT	58	0.0029

Table 6.21. full.or.part.time.employment.stat frequency and proportion table.

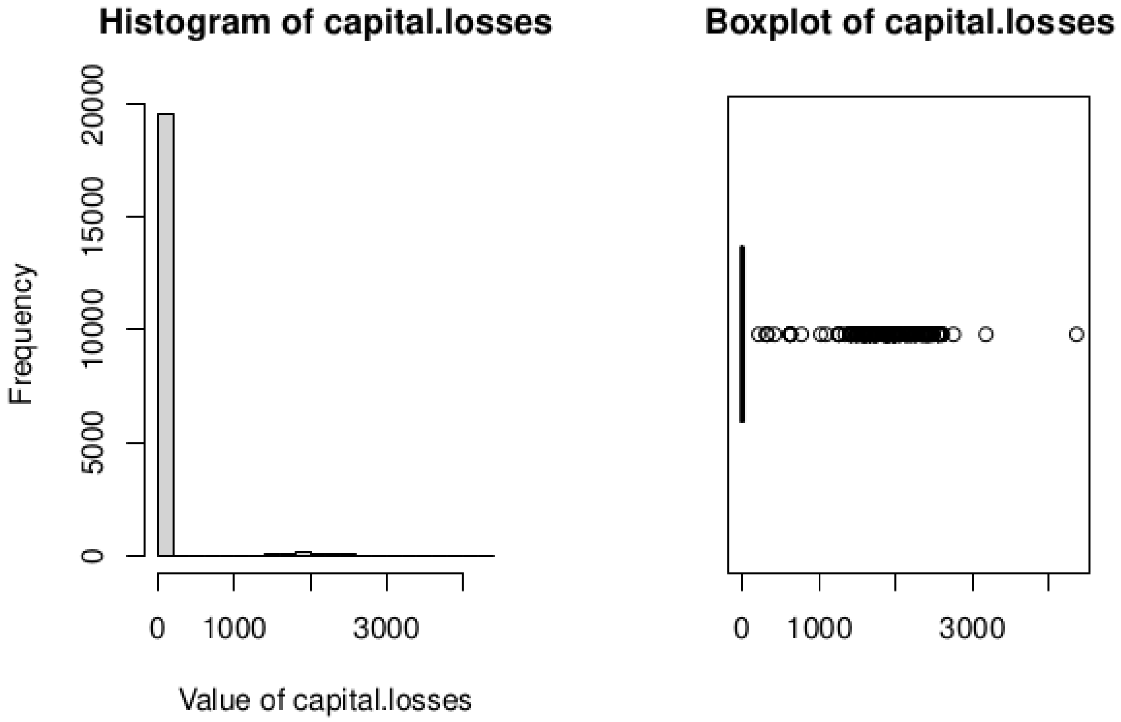
Variable 14 : capital.gains



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	493.7227	0	99999	5176.206	10.48403	0

Table 6.22. capital.gains extended Summary Statistics.

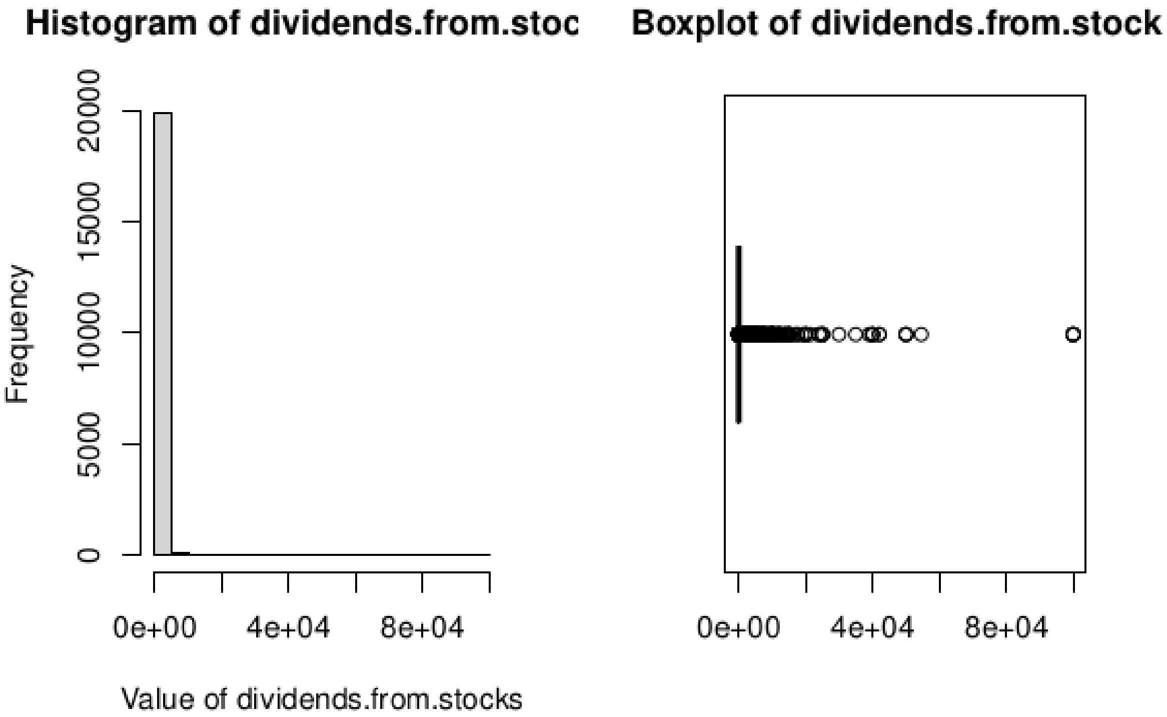
Variable 15 : capital.losses



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	36.97625	0	4356	266.4709	7.206542	0

Table 6.23. capital.losses extended Summary Statistics.

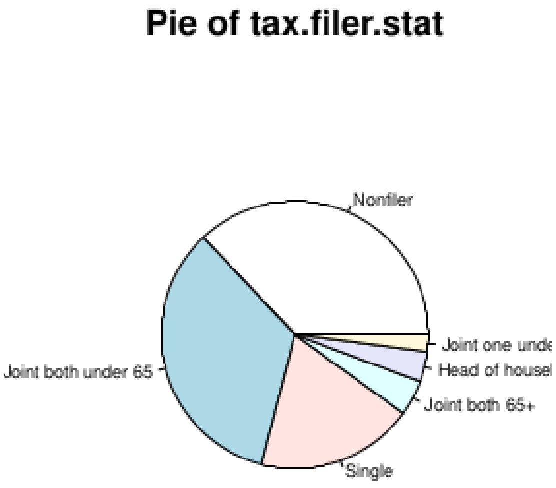
Variable 16 : dividends.from.stocks



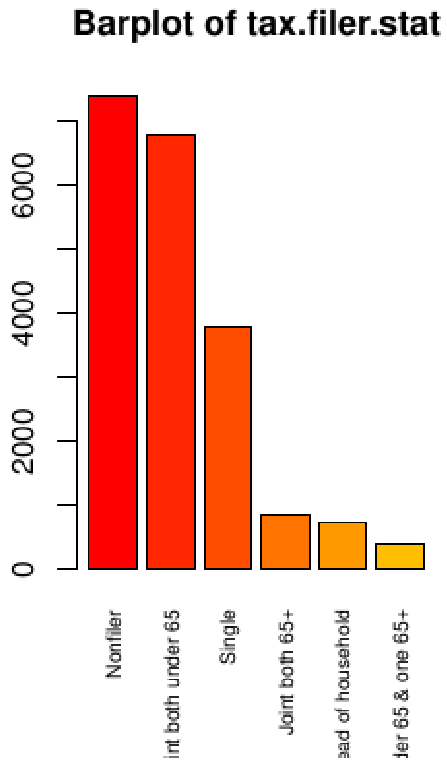
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	0	191.559	0	99999	2058.177	10.74435	0

Table 6.24. dividends.from.stocks extended Summary Statistics.

Variable 17 : tax.filer.stat



Pie of tax.filer.stat



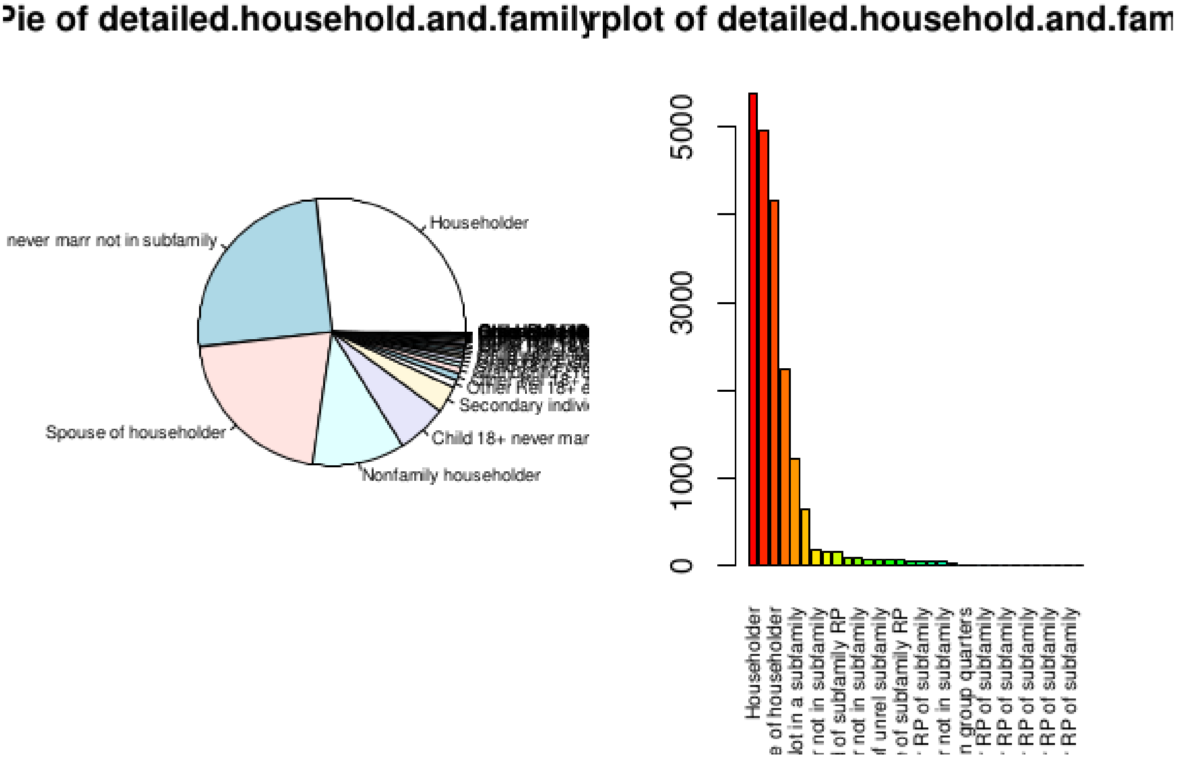
Barplot of tax.filer.stat

Number of modalities: 6

tax.filer.stat	Frequency	Proportion
Nonfiler	7401	0.37005
Joint both under 65	6802	0.34010
Single	3806	0.19030
Joint both 65+	856	0.04280
Head of household	728	0.03640
Joint one under 65 & one 65+	407	0.02035

Table 6.25. tax.filer.stat frequency and proportion table.

Variable 18 : detailed.household.and.family.stat

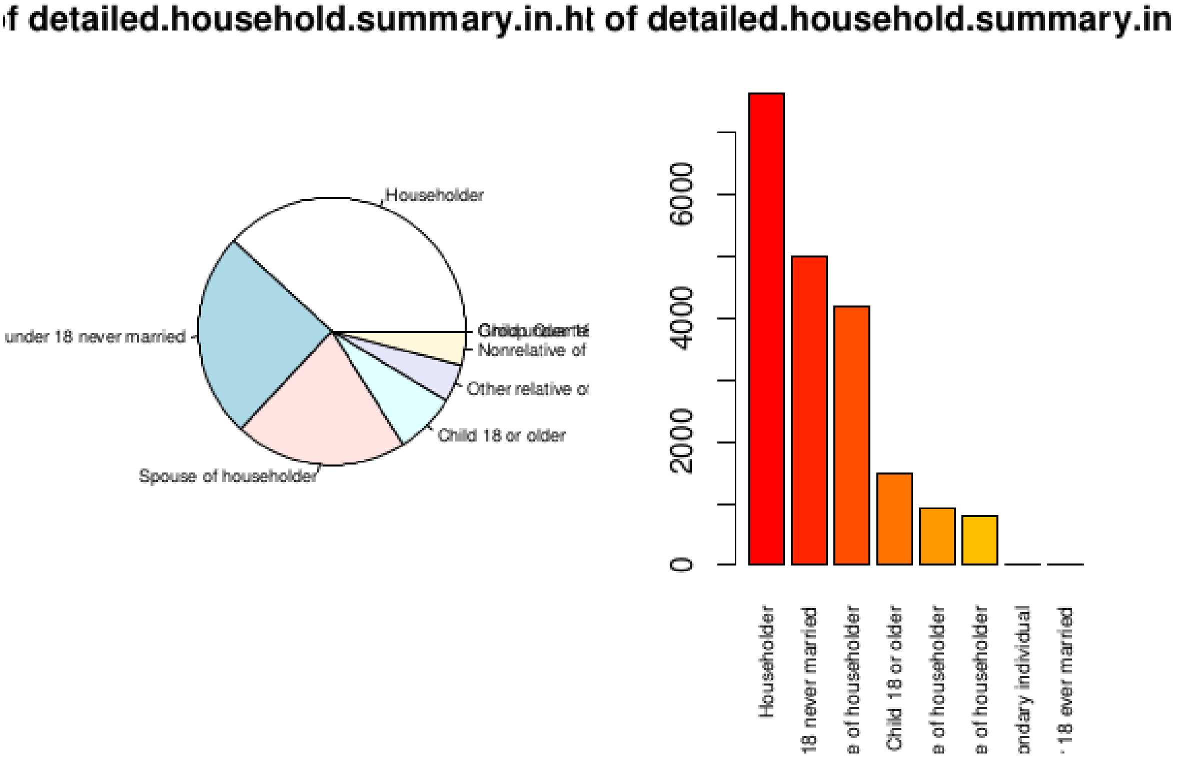


Number of modalities: 32

detailed.household.and.family.stat	Frequency	Proportion
Householder	5391	0.26955
Child <18 never marr not in subfamily	4969	0.24845
Spouse of householder	4168	0.20840
Nonfamily householder	2242	0.11210
Child 18+ never marr Not in a subfamily	1221	0.06105
Secondary individual	648	0.03240
Other Rel 18+ ever marr not in subfamily	185	0.00925
Other Rel 18+ never marr not in subfamily	168	0.00840
Grandchild <18 never marr child of subfamily RP	165	0.00825
Child 18+ ever marr Not in a subfamily	105	0.00525
Grandchild <18 never marr not in subfamily	103	0.00515
Child 18+ ever marr RP of subfamily	83	0.00415
Child under 18 of RP of unrel subfamily	80	0.00400
RP of unrelated subfamily	68	0.00340
Other Rel 18+ spouse of subfamily RP	66	0.00330
Other Rel 18+ ever marr RP of subfamily	61	0.00305
Child 18+ never marr RP of subfamily	59	0.00295
Other Rel <18 never marr child of subfamily RP	57	0.00285
Other Rel <18 never marr not in subfamily	54	0.00270
Grandchild 18+ never marr not in subfamily	43	0.00215
In group quarters	14	0.00070
Child 18+ spouse of subfamily RP	13	0.00065
Child <18 never marr RP of subfamily	10	0.00050
Spouse of RP of unrelated subfamily	8	0.00040
Other Rel 18+ never marr RP of subfamily	6	0.00030
Grandchild 18+ ever marr not in subfamily	3	0.00015
Grandchild 18+ ever marr RP of subfamily	3	0.00015
Child <18 ever marr not in subfamily	2	0.00010
Child <18 ever marr RP of subfamily	2	0.00010
Grandchild 18+ spouse of subfamily RP	1	0.00005
Other Rel <18 ever marr RP of subfamily	1	0.00005
Other Rel <18 never married RP of subfamily	1	0.00005

Table 6.26. detailed.household.and.family.stat frequency and proportion table.

Variable 19 : detailed.household.summary.in.household

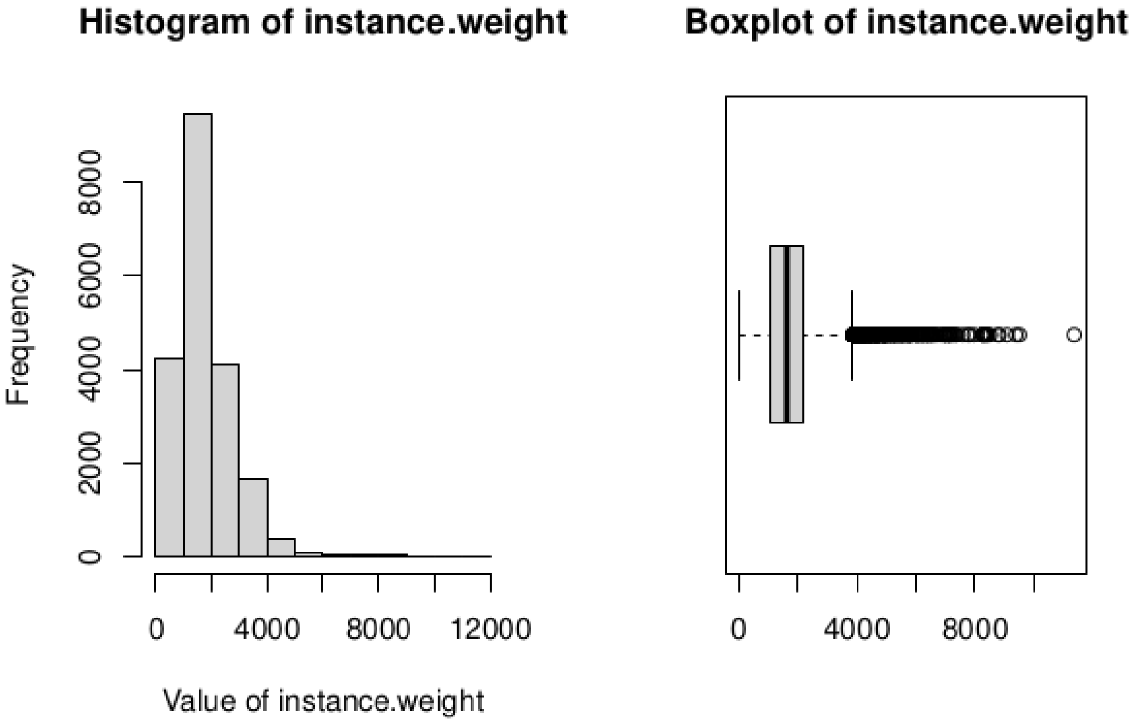


Number of modalities: 8

detailed.household.summary.in.household	Frequency	Proportion
Householder	7633	0.38165
Child under 18 never married	4981	0.24905
Spouse of householder	4170	0.20850
Child 18 or older	1481	0.07405
Other relative of householder	917	0.04585
Nonrelative of householder	806	0.04030
Group Quarters- Secondary individual	8	0.00040
Child under 18 ever married	4	0.00020

Table 6.27. detailed.household.summary.in.household frequency and proportion table.

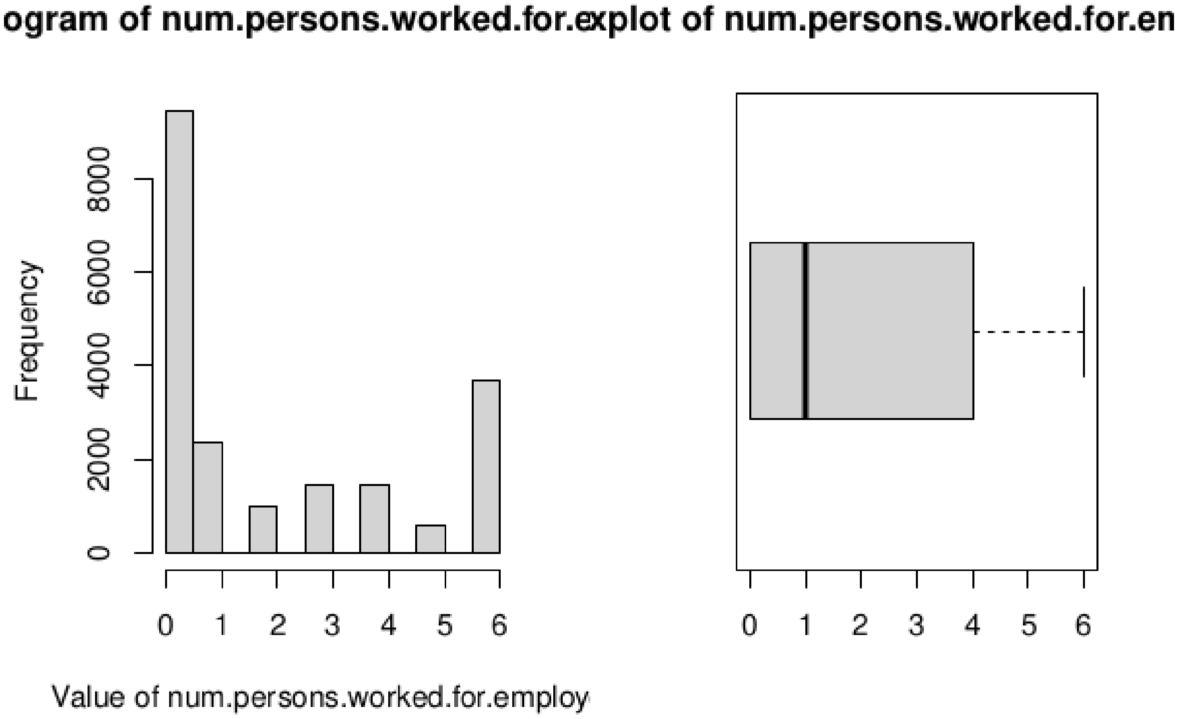
Variable 20 : instance.weight



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
40.67	1070.797	1627.42	1749.826	2187.207	11352.5	995.5348	0.5689336	0

Table 6.28. instance.weight extended Summary Statistics.

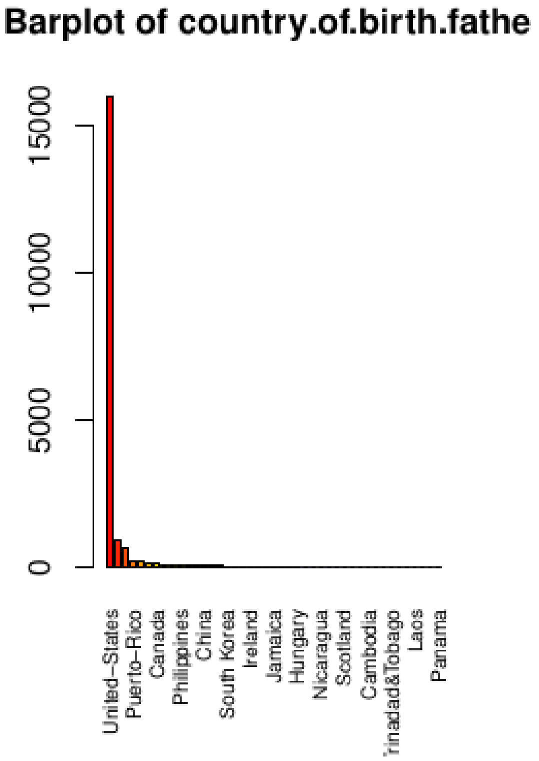
Variable 21 : num.persons.worked.for.employer



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	1	1.97915	4	6	2.365205	1.195061	0

Table 6.29. num.persons.worked.for.employer extended Summary Statistics.

Variable 22 : country.of.birth.father

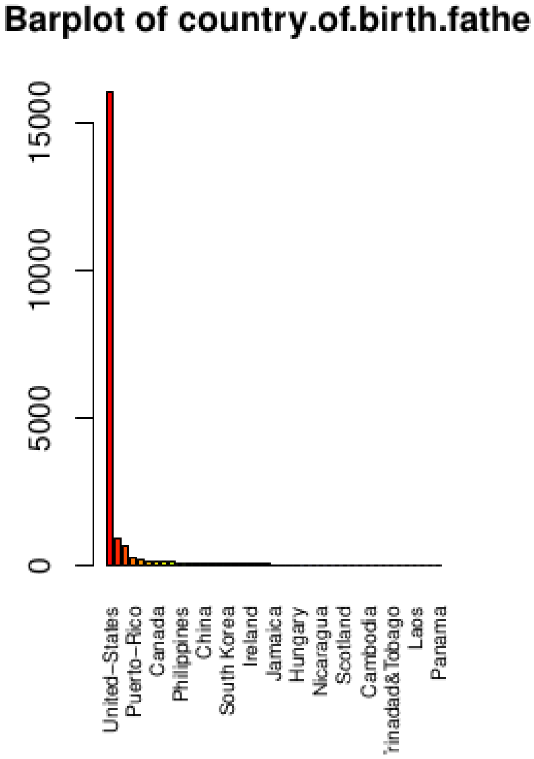


Number of modalities: 43

country.of.birth.father	Frequency	Proportion
United-States	16014	0.80070
Mexico	953	0.04765
NA	685	0.03425
Puerto-Rico	258	0.01290
Italy	208	0.01040
Germany	142	0.00710
Canada	132	0.00660
Poland	126	0.00630
Dominican-Republic	120	0.00600
Philippines	110	0.00550
Cuba	109	0.00545
El-Salvador	106	0.00530
China	85	0.00425
England	80	0.00400
Guatemala	65	0.00325
South Korea	63	0.00315
Columbia	61	0.00305
India	56	0.00280
Ireland	51	0.00255
Vietnam	51	0.00255
Japan	48	0.00240
Jamaica	43	0.00215
Portugal	41	0.00205
Haiti	33	0.00165
Hungary	33	0.00165
Peru	32	0.00160
Ecuador	29	0.00145
Nicaragua	28	0.00140
Greece	27	0.00135
Iran	25	0.00125
Scotland	23	0.00115
Yugoslavia	21	0.00105
Taiwan	20	0.00100
Cambodia	19	0.00095
France	18	0.00090
Outlying-U S (Guam USVI etc)	16	0.00080
Trinadad&Tobago	13	0.00065
Honduras	12	0.00060
Hong Kong	12	0.00060
Laos	12	0.00060
Thailand	11	0.00055
Holand-Netherlands	7	0.00035
Panama	2	0.00010

Table 6.30. country.of.birth.father frequency and proportion table.

Variable 22 : country.of.birth.father (CHANGED in preprocessing)

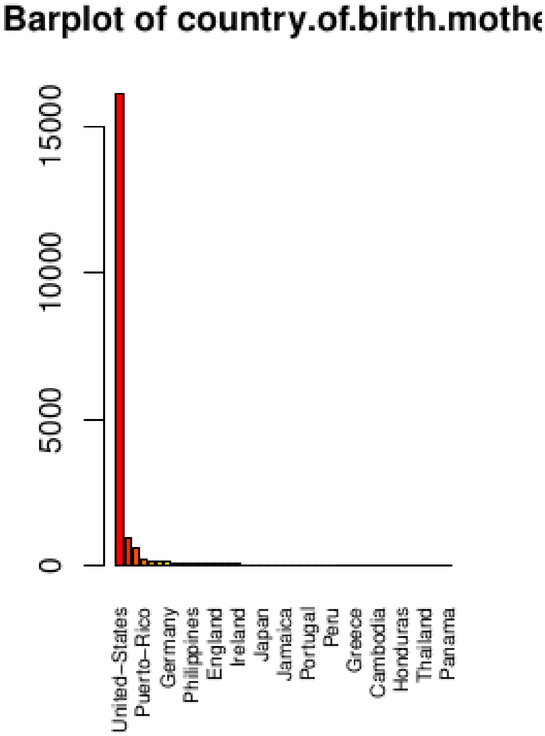


Number of modalities: 43

country.of.birth.father	Frequency	Proportion
United-States	16014	0.80070
Mexico	953	0.04765
UnknownFatherCountry	685	0.03425
Puerto-Rico	258	0.01290
Italy	208	0.01040
Germany	142	0.00710
Canada	132	0.00660
Poland	126	0.00630
Dominican-Republic	120	0.00600
Philippines	110	0.00550
Cuba	109	0.00545
El-Salvador	106	0.00530
China	85	0.00425
England	80	0.00400
Guatemala	65	0.00325
South Korea	63	0.00315
Columbia	61	0.00305
India	56	0.00280
Ireland	51	0.00255
Vietnam	51	0.00255
Japan	48	0.00240
Jamaica	43	0.00215
Portugal	41	0.00205
Haiti	33	0.00165
Hungary	33	0.00165
Peru	32	0.00160
Ecuador	29	0.00145
Nicaragua	28	0.00140
Greece	27	0.00135
Iran	25	0.00125
Scotland	23	0.00115
Yugoslavia	21	0.00105
Taiwan	20	0.00100
Cambodia	19	0.00095
France	18	0.00090
Outlying-U S (Guam USVI etc)	16	0.00080
Trinadad&Tobago	13	0.00065
Honduras	12	0.00060
Hong Kong	12	0.00060
Laos	12	0.00060
Thailand	11	0.00055
Holand-Netherlands	7	0.00035
Panama	2	0.00010

Table 6.31. country.of.birth.father frequency and proportion table.

Variable 23 : country.of.birth.mother

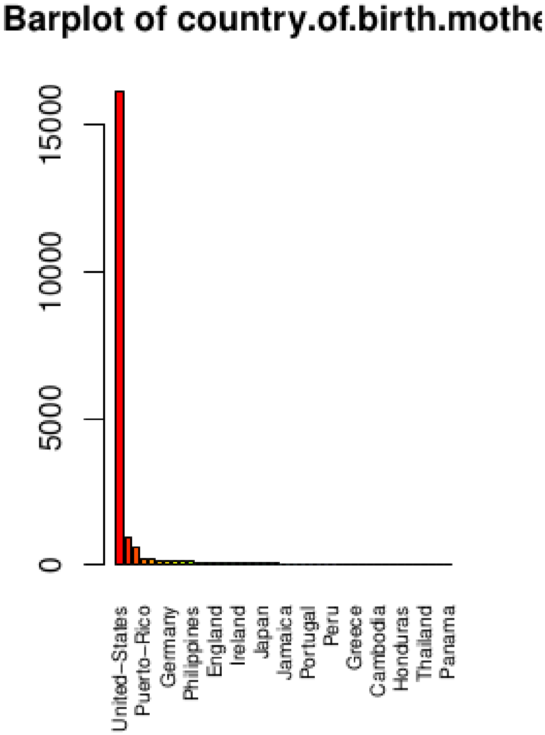


Number of modalities: 43

country.of.birth.mother	Frequency	Proportion
United-States	16129	0.80645
Mexico	939	0.04695
NA	597	0.02985
Puerto-Rico	237	0.01185
Italy	179	0.00895
Canada	150	0.00750
Germany	143	0.00715
Poland	124	0.00620
El-Salvador	114	0.00570
Philippines	112	0.00560
Cuba	106	0.00530
Dominican-Republic	104	0.00520
England	104	0.00520
South Korea	77	0.00385
China	74	0.00370
Ireland	67	0.00335
Guatemala	63	0.00315
Columbia	62	0.00310
Japan	59	0.00295
India	58	0.00290
Vietnam	49	0.00245
Jamaica	41	0.00205
Haiti	34	0.00170
Hungary	33	0.00165
Portugal	33	0.00165
Ecuador	32	0.00160
Nicaragua	28	0.00140
Peru	28	0.00140
France	24	0.00120
Taiwan	22	0.00110
Greece	21	0.00105
Scotland	19	0.00095
Iran	18	0.00090
Cambodia	17	0.00085
Outlying-U S (Guam USVI etc)	17	0.00085
Yugoslavia	17	0.00085
Honduras	16	0.00080
Hong Kong	13	0.00065
Laos	12	0.00060
Thailand	11	0.00055
Trinadad&Tobago	9	0.00045
Holand-Netherlands	7	0.00035
Panama	1	0.00005

Table 6.32. country.of.birth.mother frequency and proportion table.

Variable 23 : country.of.birth.mother (CHANGED in preprocessing)

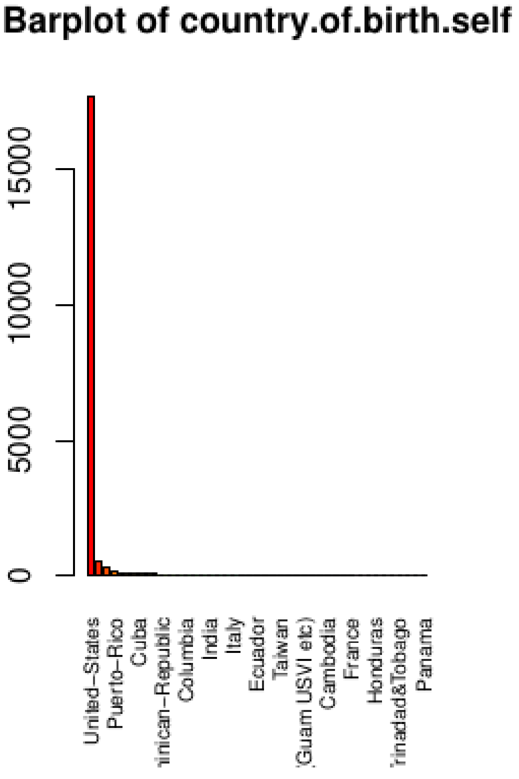


Number of modalities: 43

country.of.birth.mother	Frequency	Proportion
United-States	16129	0.80645
Mexico	939	0.04695
UnknownMotherCountry	597	0.02985
Puerto-Rico	237	0.01185
Italy	179	0.00895
Canada	150	0.00750
Germany	143	0.00715
Poland	124	0.00620
El-Salvador	114	0.00570
Philippines	112	0.00560
Cuba	106	0.00530
Dominican-Republic	104	0.00520
England	104	0.00520
South Korea	77	0.00385
China	74	0.00370
Ireland	67	0.00335
Guatemala	63	0.00315
Columbia	62	0.00310
Japan	59	0.00295
India	58	0.00290
Vietnam	49	0.00245
Jamaica	41	0.00205
Haiti	34	0.00170
Hungary	33	0.00165
Portugal	33	0.00165
Ecuador	32	0.00160
Nicaragua	28	0.00140
Peru	28	0.00140
France	24	0.00120
Taiwan	22	0.00110
Greece	21	0.00105
Scotland	19	0.00095
Iran	18	0.00090
Cambodia	17	0.00085
Outlying-U S (Guam USVI etc)	17	0.00085
Yugoslavia	17	0.00085
Honduras	16	0.00080
Hong Kong	13	0.00065
Laos	12	0.00060
Thailand	11	0.00055
Trinadad&Tobago	9	0.00045
Holand-Netherlands	7	0.00035
Panama	1	0.00005

Table 6.33. country.of.birth.mother frequency and proportion table.

Variable 24 : country.of.birth.self

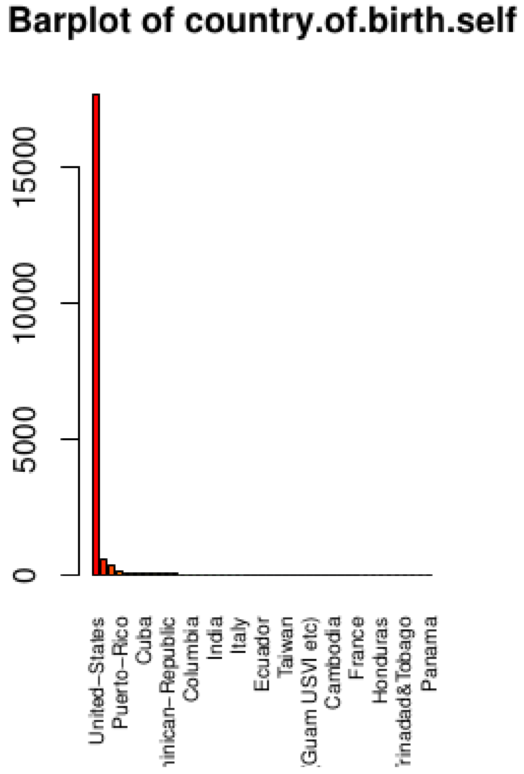


Number of modalities: 43

country.of.birth.self	Frequency	Proportion
United-States	17686	0.88430
Mexico	558	0.02790
NA	359	0.01795
Puerto-Rico	151	0.00755
Philippines	90	0.00450
Germany	84	0.00420
Cuba	81	0.00405
Canada	78	0.00390
El-Salvador	77	0.00385
Dominican-Republic	62	0.00310
South Korea	54	0.00270
Poland	49	0.00245
Columbia	46	0.00230
England	46	0.00230
Guatemala	46	0.00230
India	45	0.00225
China	44	0.00220
Japan	44	0.00220
Italy	41	0.00205
Vietnam	41	0.00205
Jamaica	29	0.00145
Ecuador	25	0.00125
Peru	24	0.00120
Haiti	23	0.00115
Taiwan	23	0.00115
Ireland	19	0.00095
Nicaragua	19	0.00095
Outlying-U S (Guam USVI etc)	17	0.00085
Iran	15	0.00075
Portugal	15	0.00075
Cambodia	14	0.00070
Greece	14	0.00070
Laos	11	0.00055
France	10	0.00050
Scotland	10	0.00050
Thailand	10	0.00050
Honduras	9	0.00045
Hong Kong	9	0.00045
Hungary	7	0.00035
Trinadad&Tobago	6	0.00030
Yugoslavia	5	0.00025
Holand-Netherlands	3	0.00015
Panama	1	0.00005

Table 6.34. country.of.birth.self frequency and proportion table.

Variable 24 : country.of.birth.self (CHANGED in preprocessing)

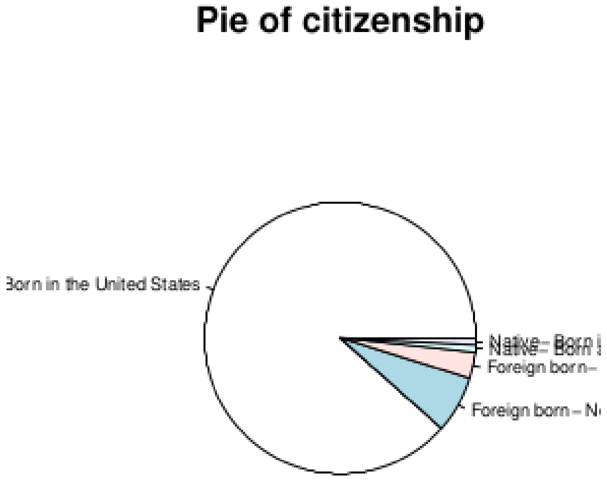


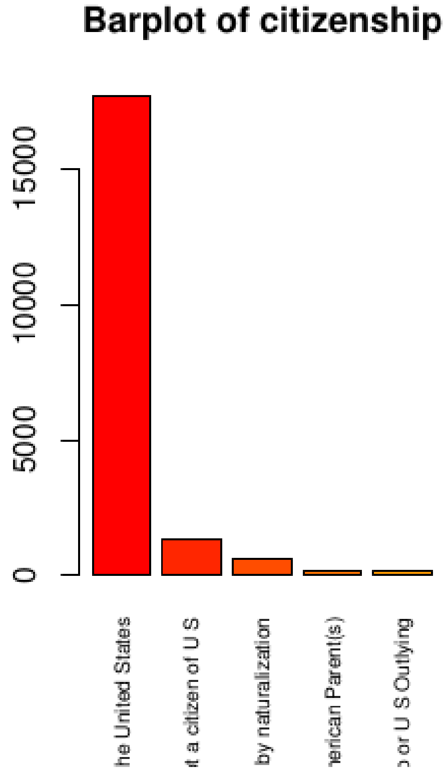
Number of modalities: 43

country.of.birth.self	Frequency	Proportion
United-States	17686	0.88430
Mexico	558	0.02790
UnknownSelfCountry	359	0.01795
Puerto-Rico	151	0.00755
Philippines	90	0.00450
Germany	84	0.00420
Cuba	81	0.00405
Canada	78	0.00390
El-Salvador	77	0.00385
Dominican-Republic	62	0.00310
South Korea	54	0.00270
Poland	49	0.00245
Columbia	46	0.00230
England	46	0.00230
Guatemala	46	0.00230
India	45	0.00225
China	44	0.00220
Japan	44	0.00220
Italy	41	0.00205
Vietnam	41	0.00205
Jamaica	29	0.00145
Ecuador	25	0.00125
Peru	24	0.00120
Haiti	23	0.00115
Taiwan	23	0.00115
Ireland	19	0.00095
Nicaragua	19	0.00095
Outlying-U S (Guam USVI etc)	17	0.00085
Iran	15	0.00075
Portugal	15	0.00075
Cambodia	14	0.00070
Greece	14	0.00070
Laos	11	0.00055
France	10	0.00050
Scotland	10	0.00050
Thailand	10	0.00050
Honduras	9	0.00045
Hong Kong	9	0.00045
Hungary	7	0.00035
Trinadad&Tobago	6	0.00030
Yugoslavia	5	0.00025
Holand-Netherlands	3	0.00015
Panama	1	0.00005

Table 6.35. country.of.birth.self frequency and proportion table.

Variable 25 : citizenship



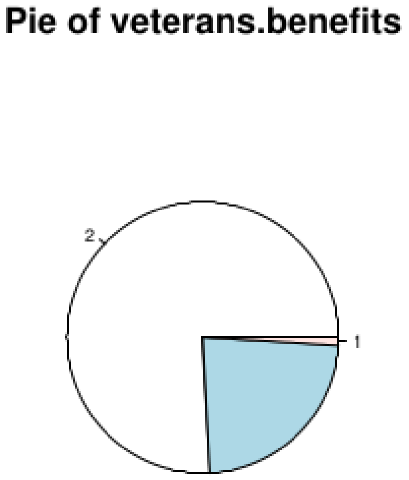


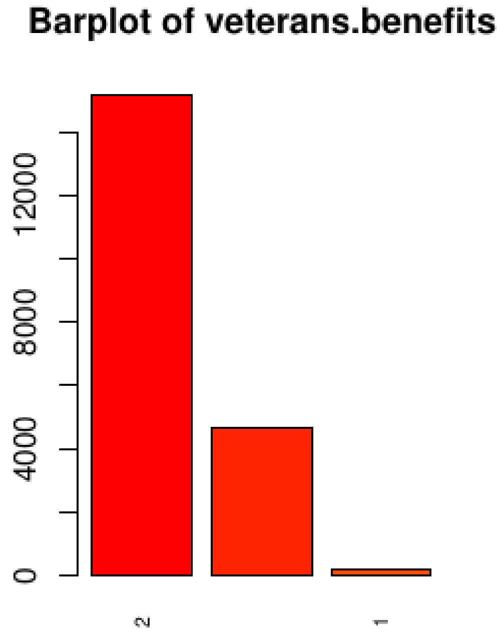
Number of modalities: 5

citizenship	Frequency	Proportion
Native- Born in the United States	17686	0.8843
Foreign born- Not a citizen of U S	1356	0.0678
Foreign born- U S citizen by naturalization	616	0.0308
Native- Born abroad of American Parent(s)	174	0.0087
Native- Born in Puerto Rico or U S Outlying	168	0.0084

Table 6.36. citizenship frequency and proportion table.

Variable 26 : veterans.benefits


Pie of veterans.benefits

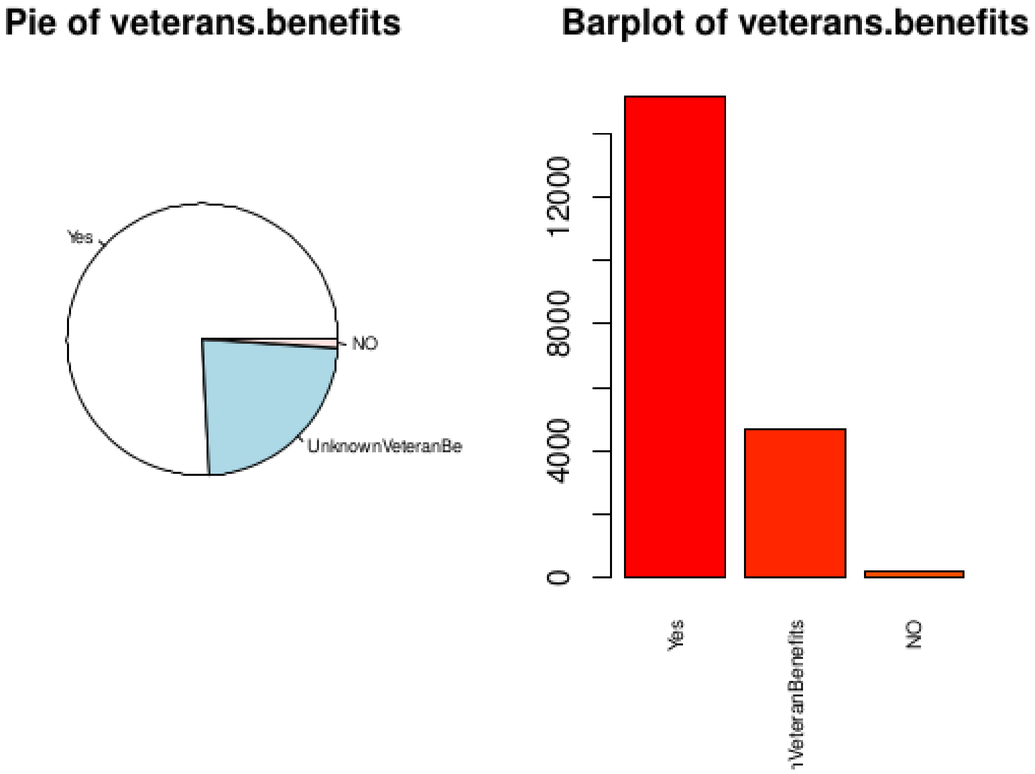

Barplot of veterans.benefits

Number of modalities: 3

veterans.benefits	Frequency	Proportion
2	15158	0.75790
NA	4651	0.23255
1	191	0.00955

Table 6.37. veterans.benefits frequency and proportion table.

Variable 26 : veterans.benefits (CHANGED in preprocessing)

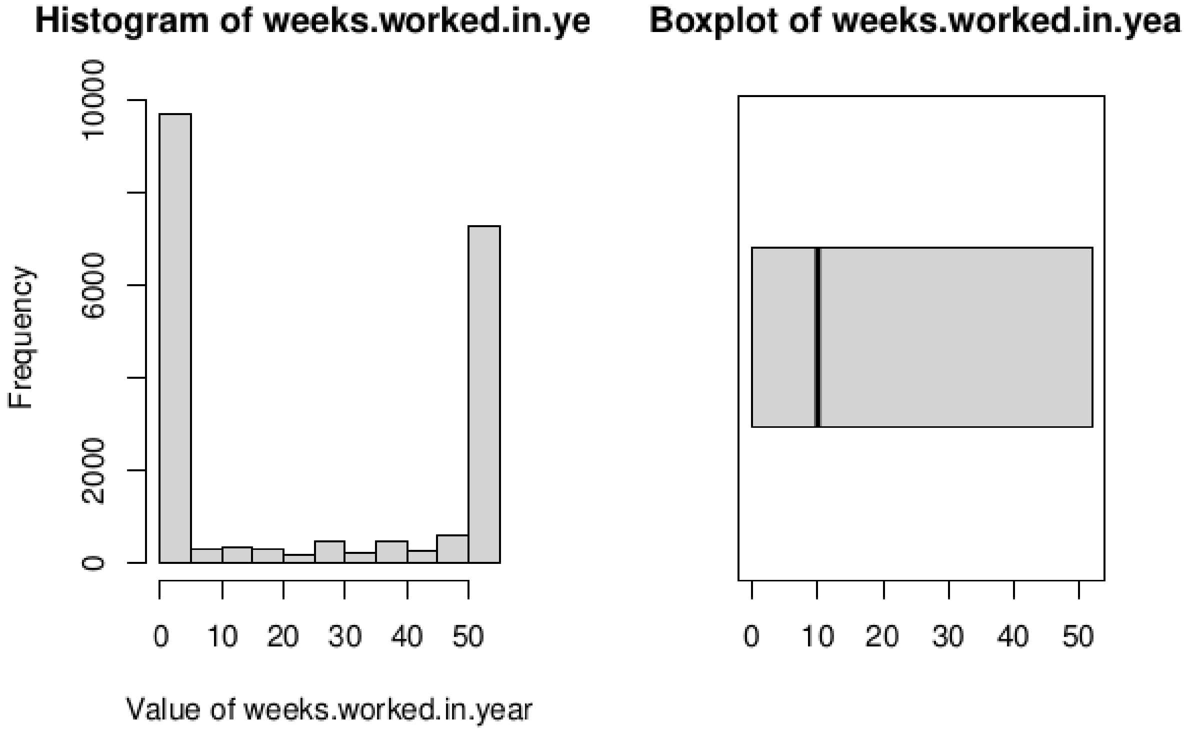


Number of modalities: 3

veterans.benefits	Frequency	Proportion
Yes	15158	0.75790
UnknownVeteranBenefits	4651	0.23255
NO	191	0.00955

Table 6.38. veterans.benefits frequency and proportion table.

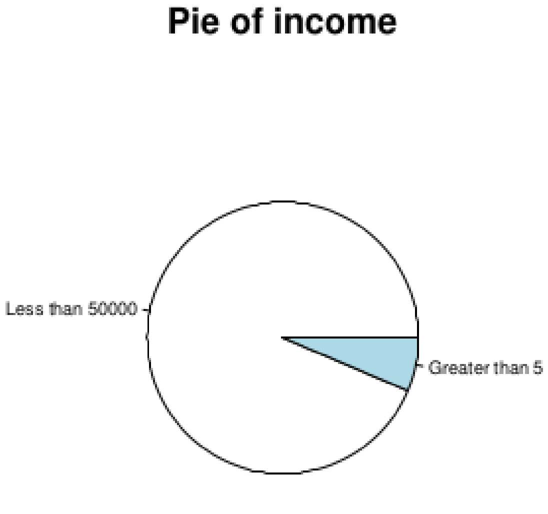
Variable 27 : weeks.worked.in.year

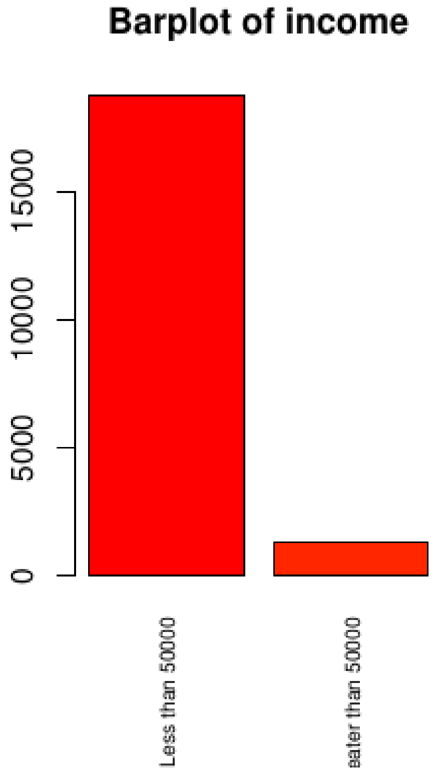


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd	vc.	Missing
0	0	10	23.5184	52	52	24.45392	1.039778	0

Table 6.39. weeks.worked.in.year extended Summary Statistics.

Variable 28 : income



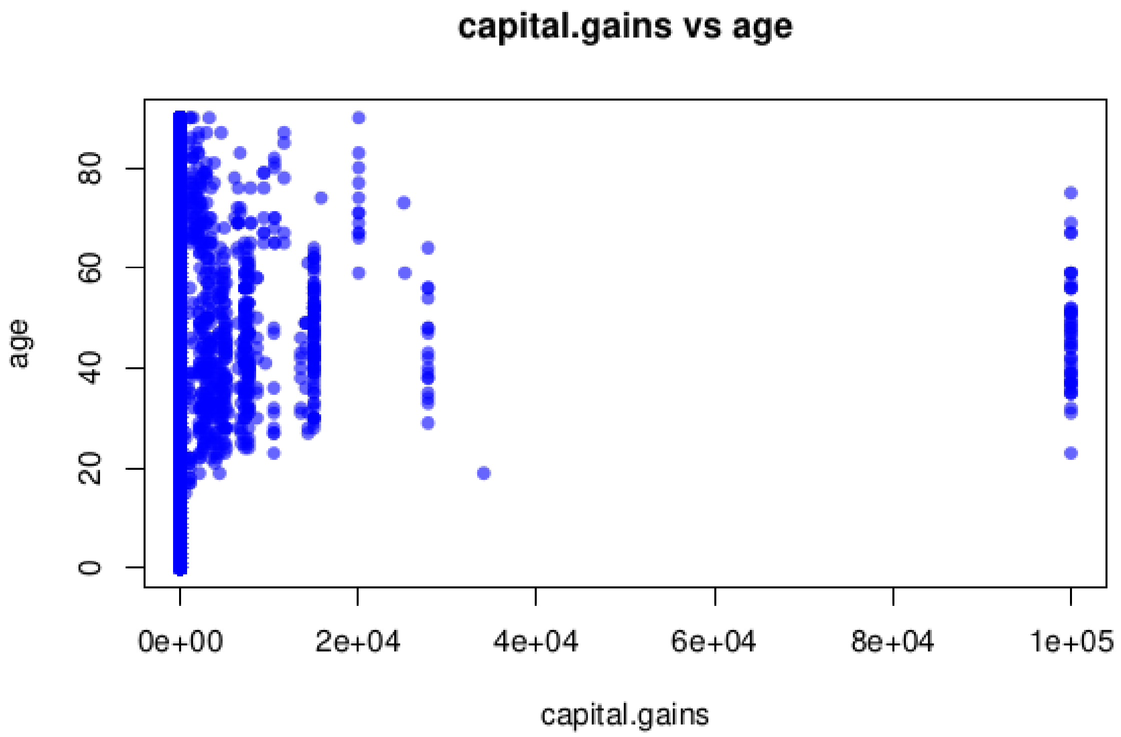


Number of modalities: 2

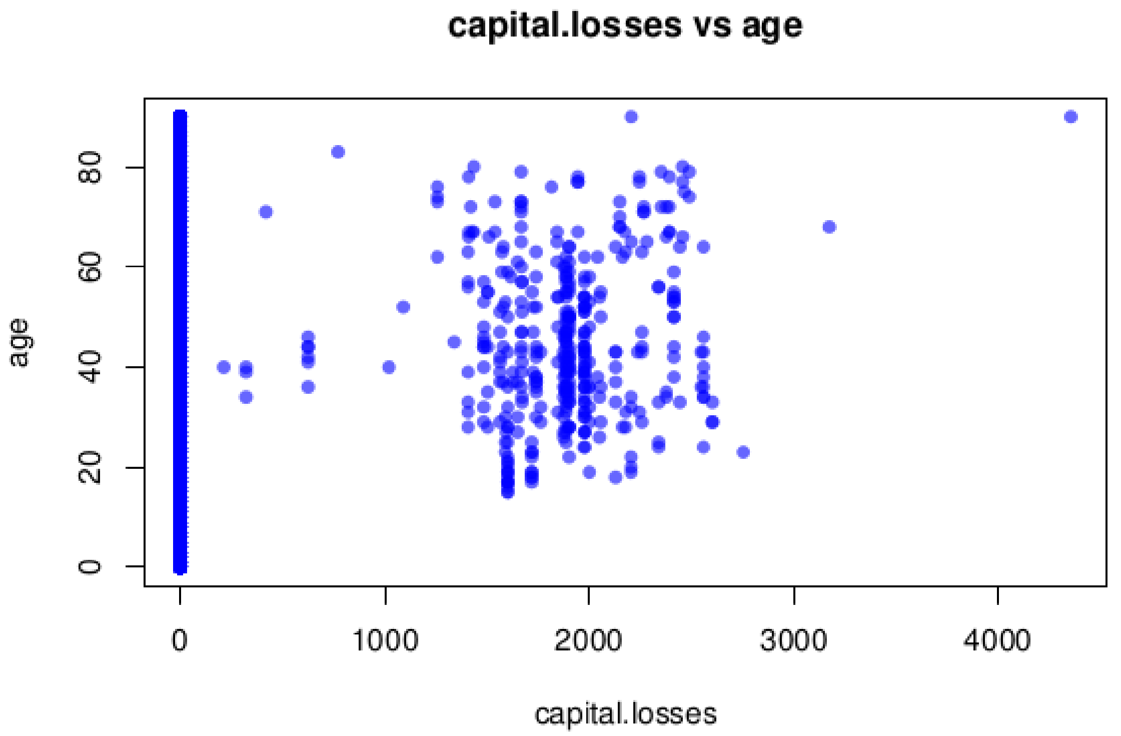
income	Frequency	Proportion
Less than 50000	18728	0.9364
Greater than 50000	1272	0.0636

Table 6.40. income frequency and proportion table.

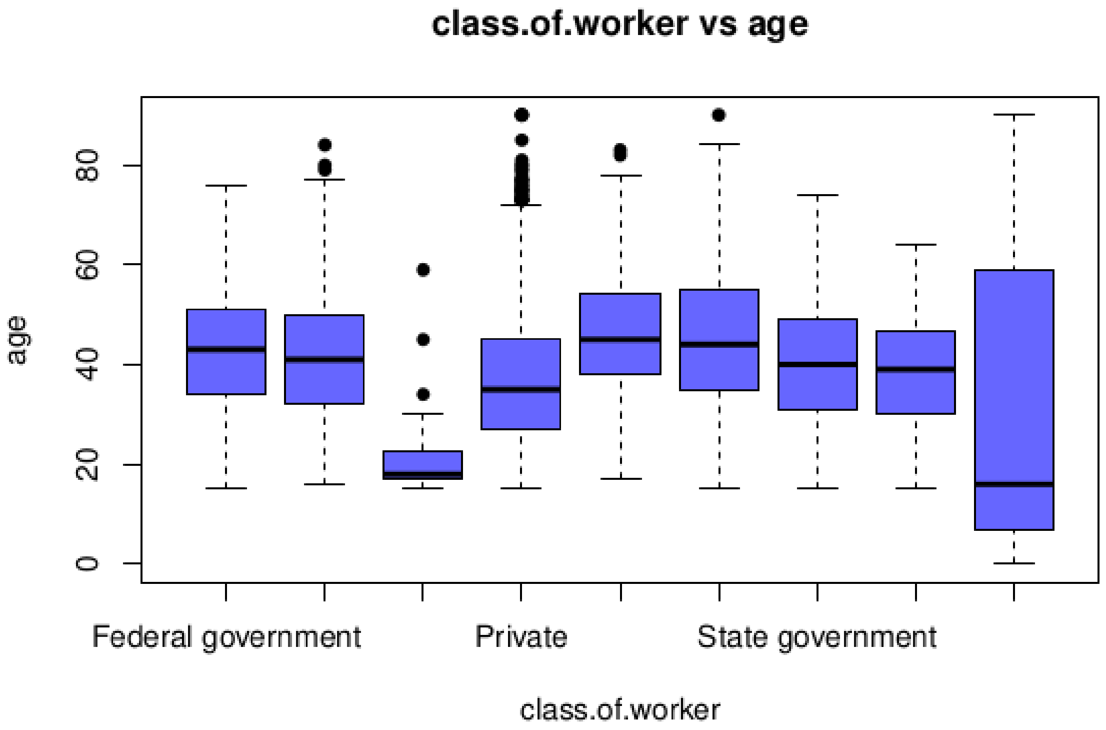
4.2 Bivariate plots

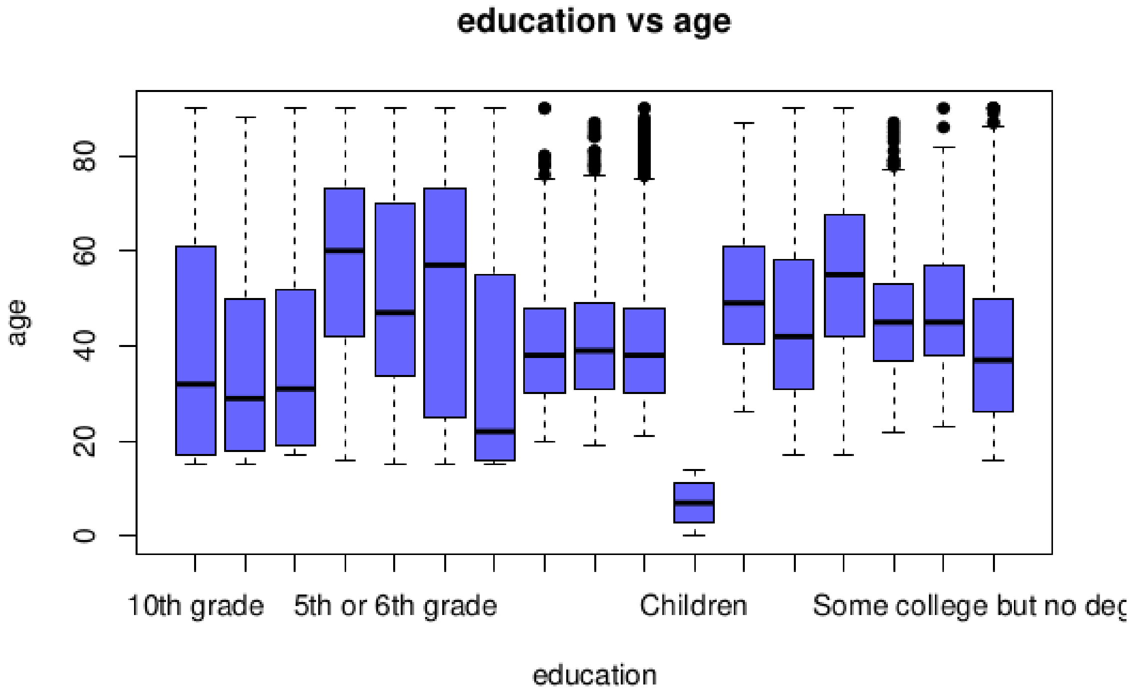


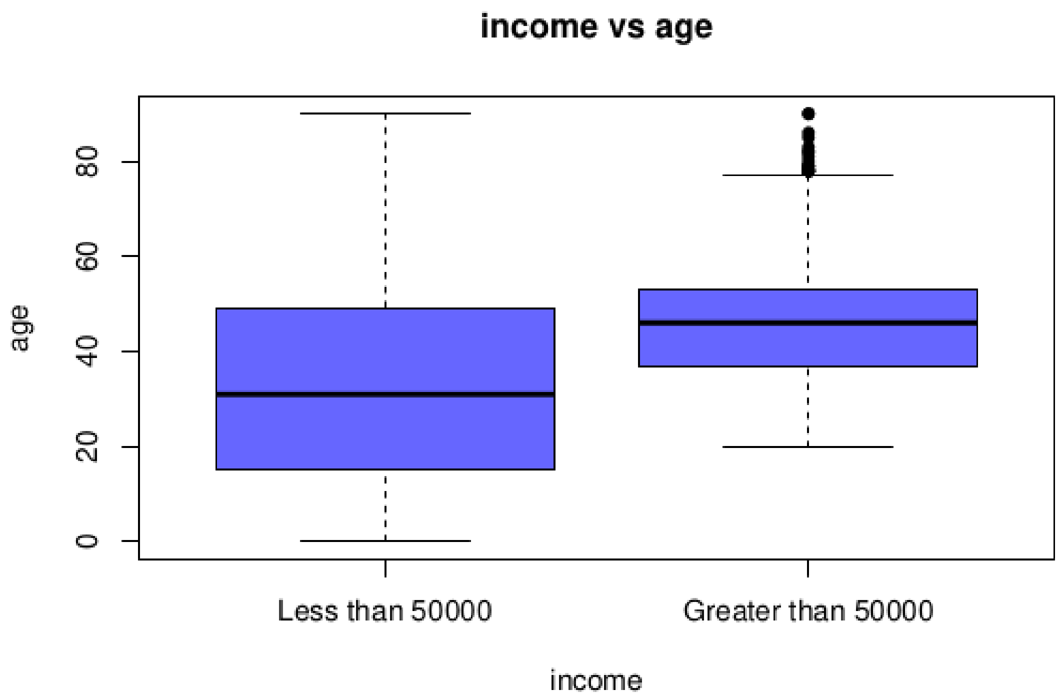
Correlation between capital.gains and age : 0.0561878766721521

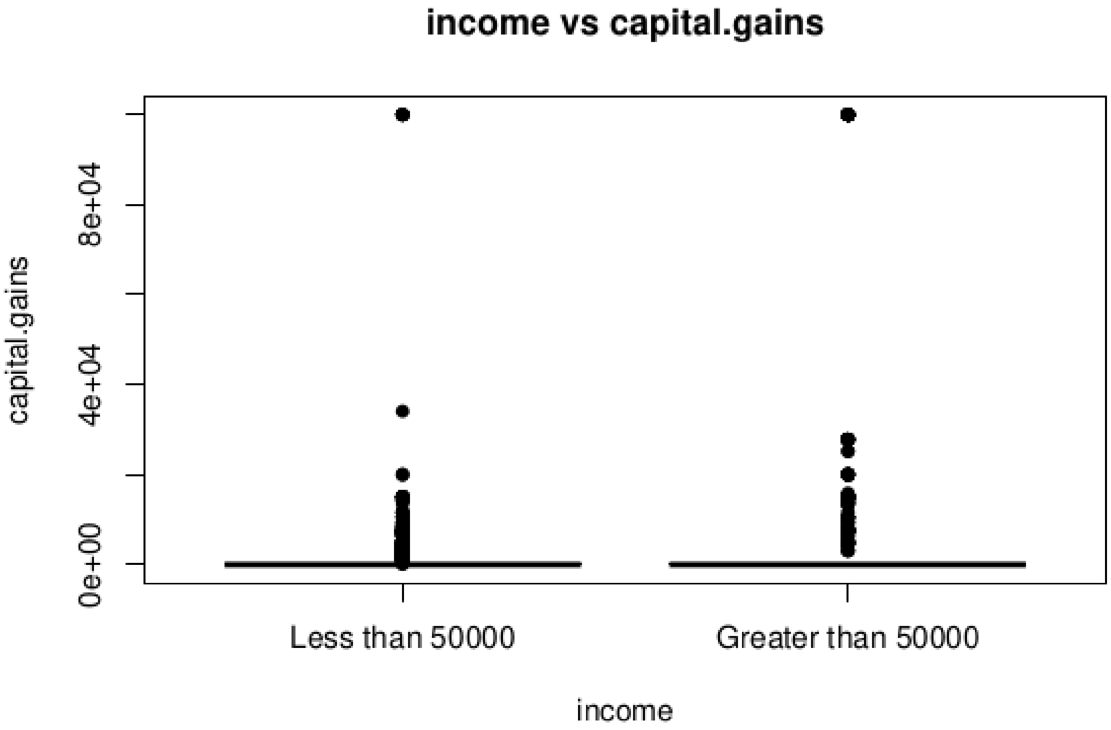


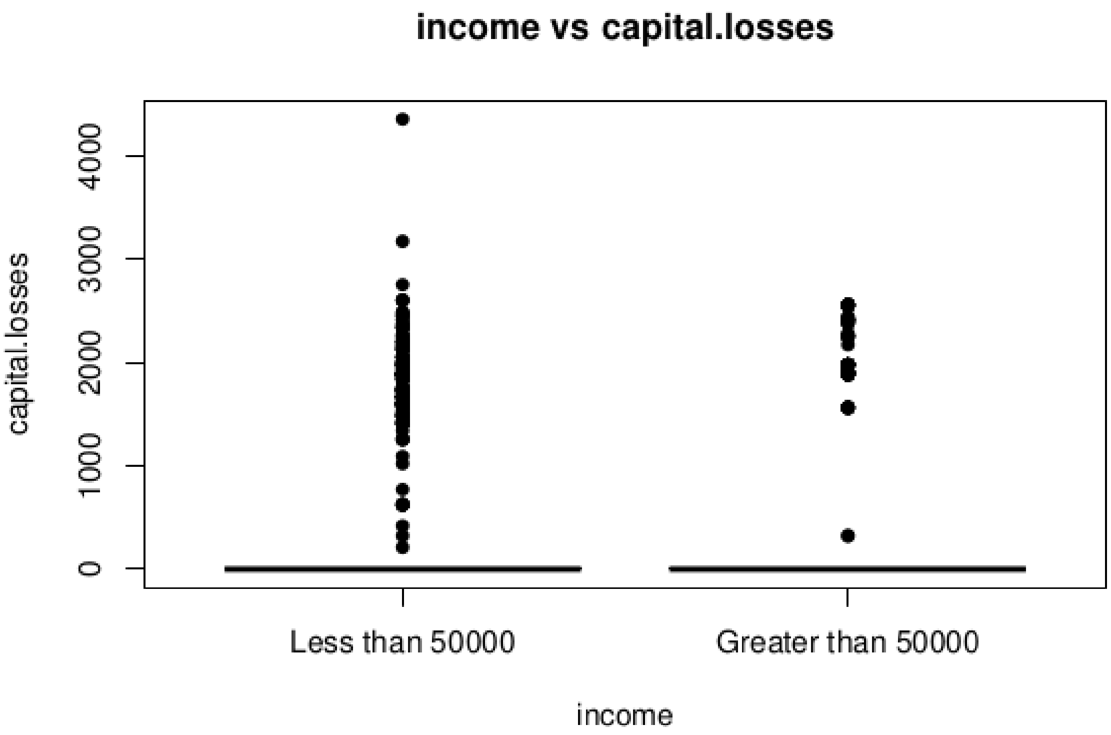
Correlation between capital.losses and age : 0.0705441910975547

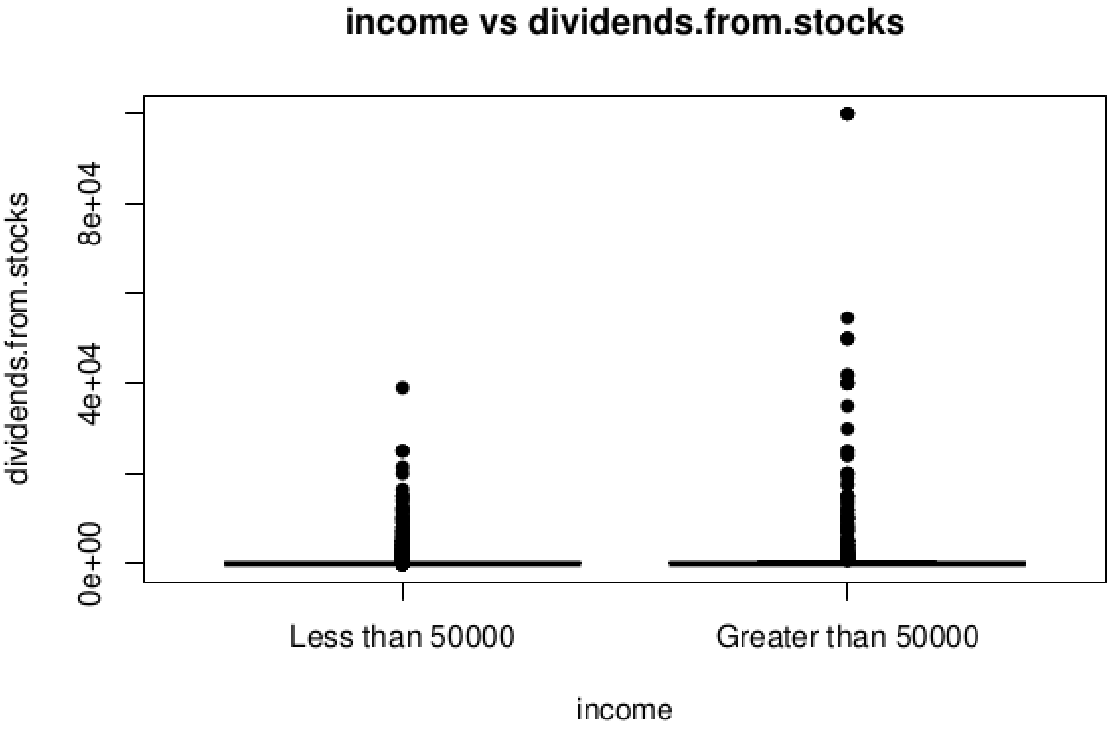

class.of.worker vs age
age
0
20
40
60
80
Federal government
Private
State government
class.of.worker

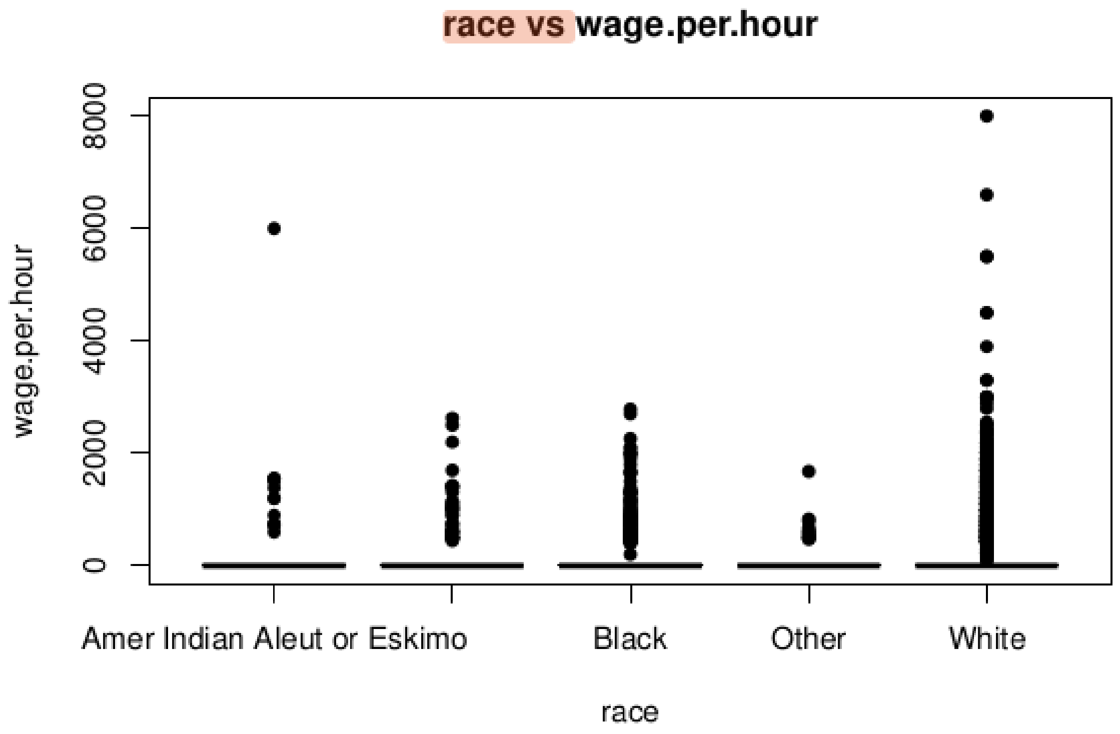

education vs age
age
0
20
40
60
80
10th grade
5th or 6th grade
Children
Some college but no deg
education

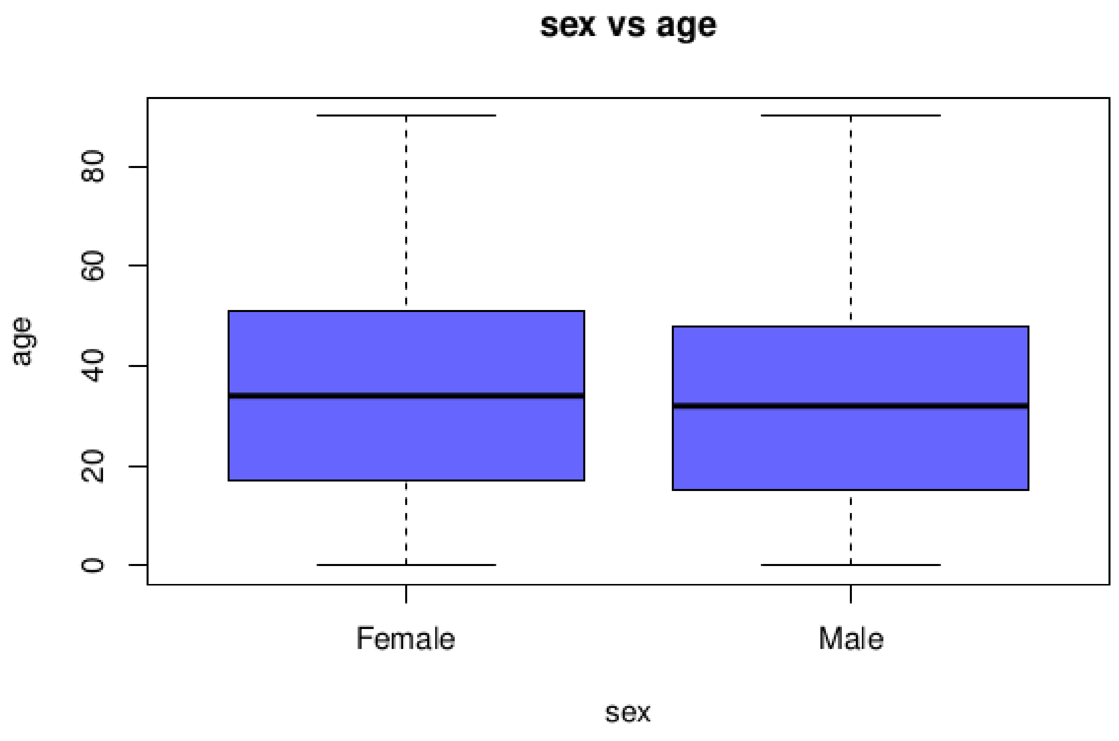

income vs age
age
0
20
40
60
80
Less than 50000
Greater than 50000
income

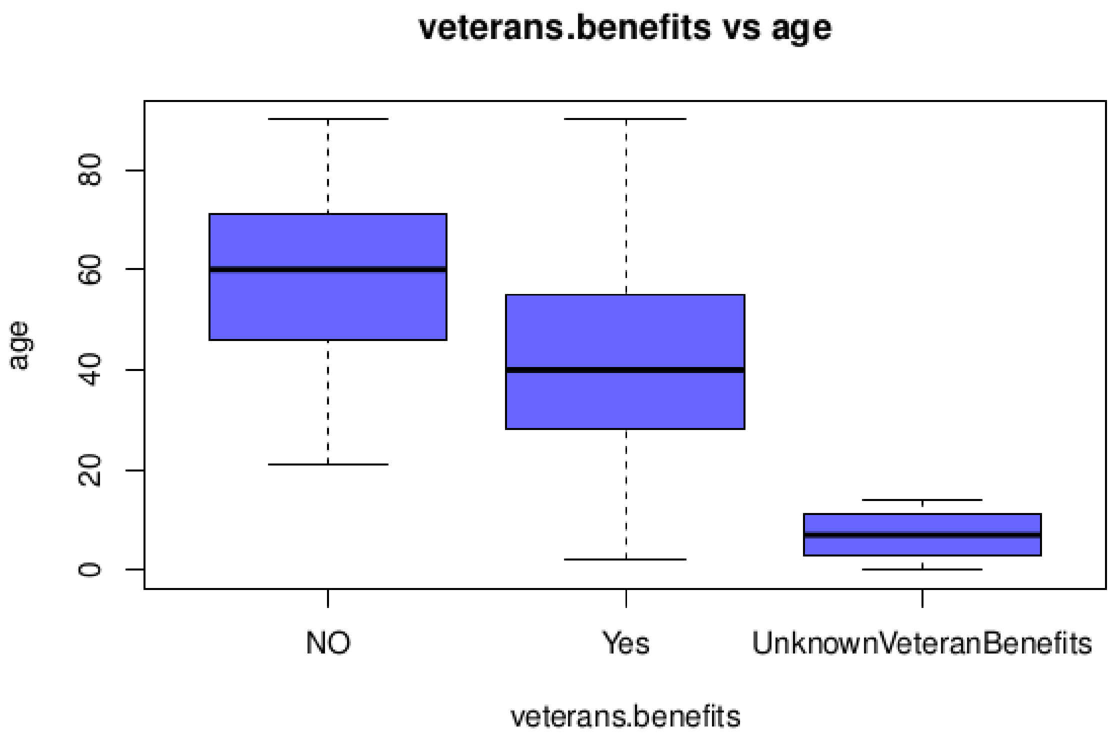

income vs capital.gains
capital.gains
0e+00
4e+04
8e+04
Less than 50000
Greater than 50000
income

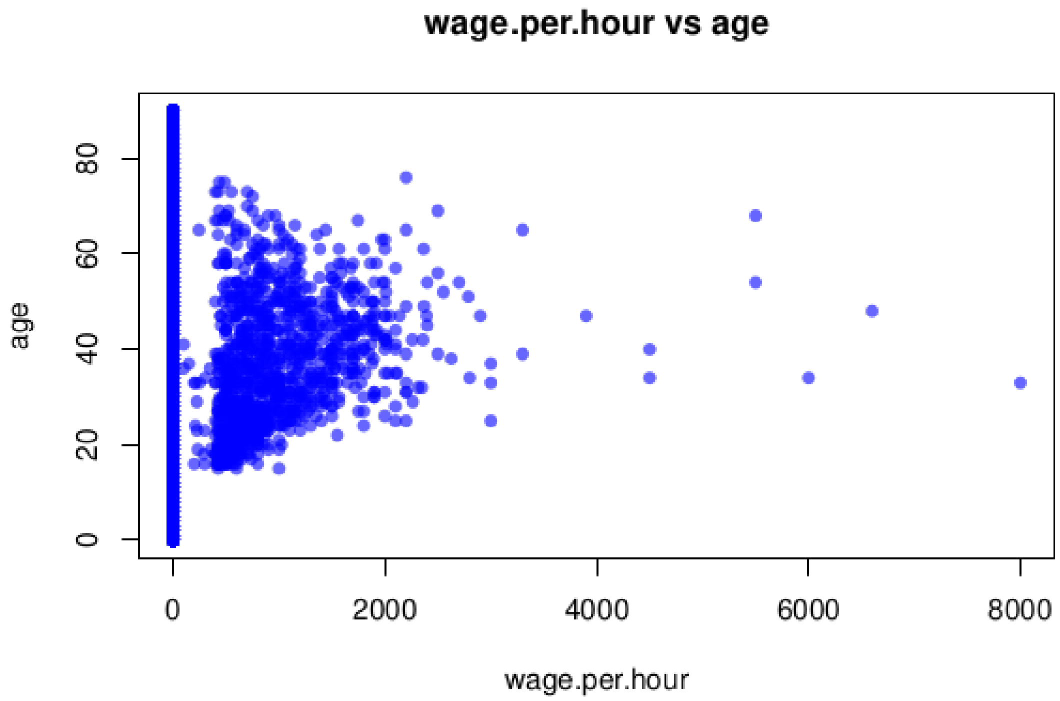

income vs capital.losses
capital.losses
0
1000
2000
3000
4000
Less than 50000
Greater than 50000
income


income vs dividends.from.stocks
dividends.from.stocks
0e+00
4e+04
8e+04
Less than 50000
Greater than 50000
income


race vs wage.per.hour
wage.per.hour
0
2000
4000
6000
8000
Amer Indian Aleut or Eskimo
Black
Other
White
race


sex vs age
age
0
20
40
60
80
Female
Male
sex





Correlation between wage.per.hour and age : 0.0409768809322131

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.

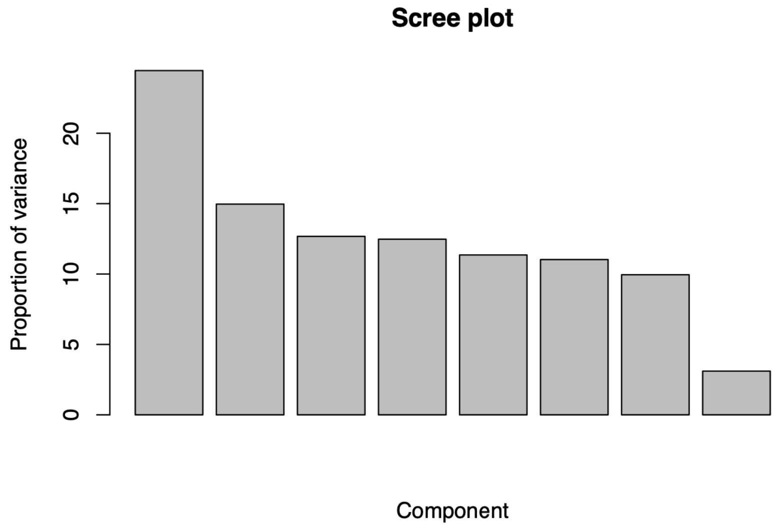
- It is also worth noting that a significant proportion (61%) are either in the armed forces or children and 21% have a full-time schedule. Finally, the majority (93%) have an income less than $50,000 while only 7% have an income greater than $50,000.

5.PCA analysis for numerical variables

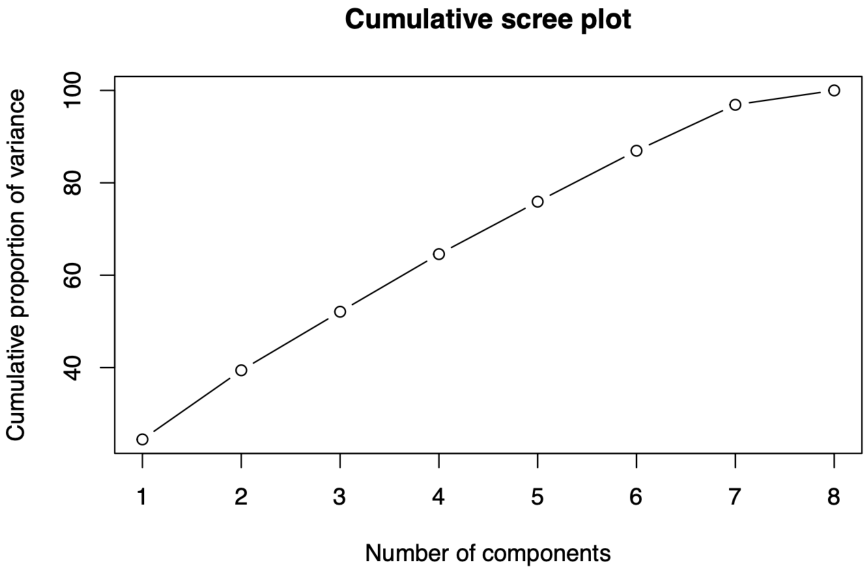
PCA is a useful technique for identifying patterns and relationships among variables and reducing the dimensionality of a dataset

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

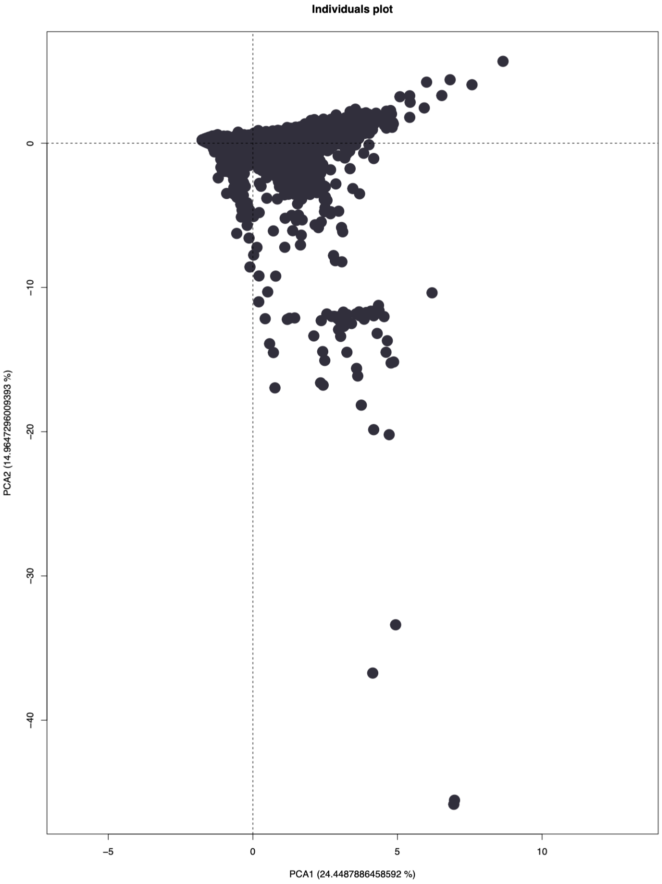
A factorial map is a graphical representation of the relationship between several variables in a dataset. It is a type of dimensionality reduction technique that is used to explore and visualize high-dimensional data. Factorial maps can be created using methods such as principal component analysis (PCA).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

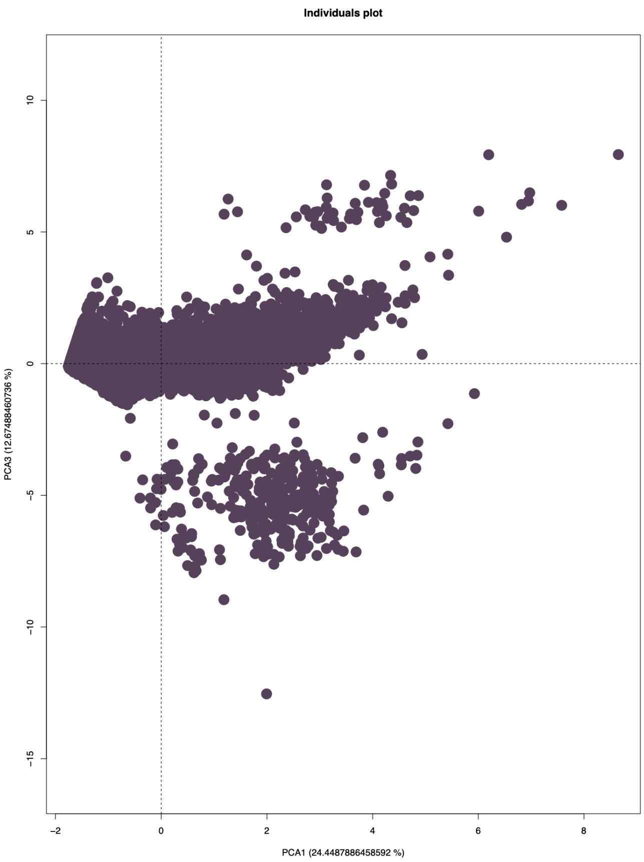
Factorial maps are useful for gaining insights into complex datasets, identifying patterns and trends, and detecting outliers or anomalies. They can also be used for exploratory data analysis, clustering, classification, and visualization of high-dimensional data.

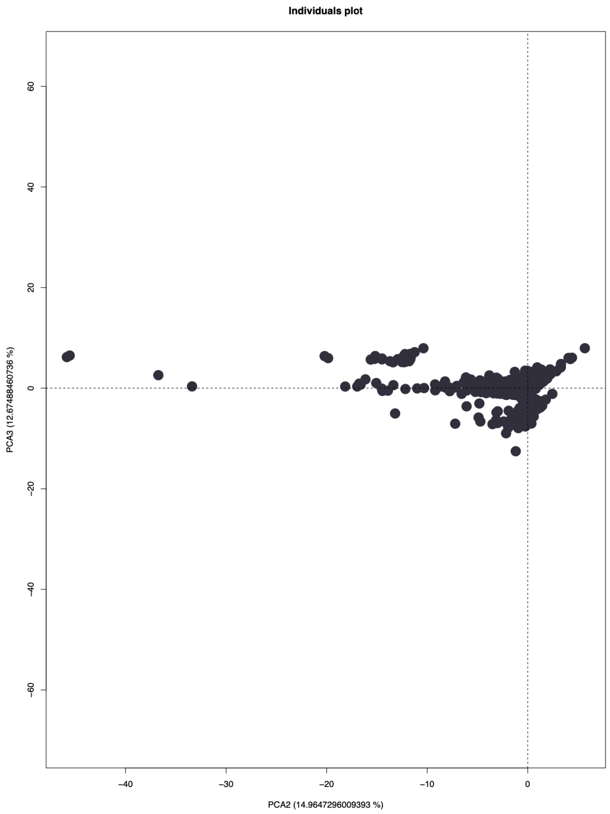
5.2.1 Individuals plot

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.

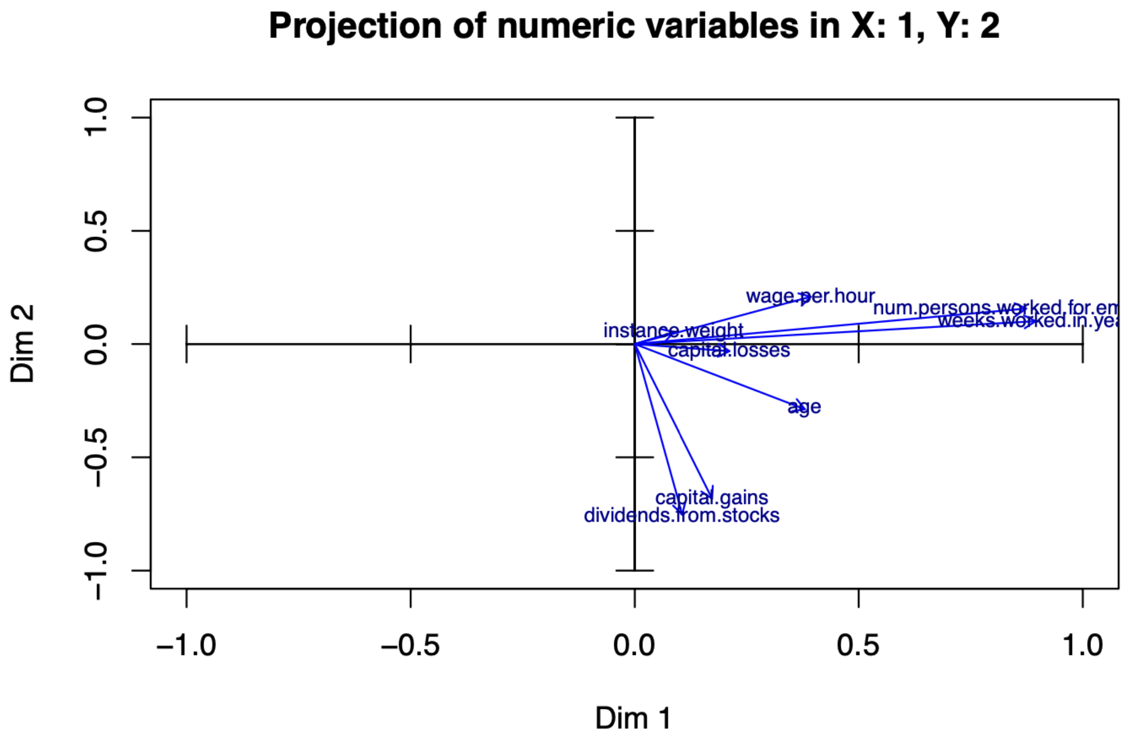



Individuals plot
PCA3 (12.67488460736 %)
10
5
0
−5
−10
−15
PCA1 (24.4487886458592 %)
−2
0
2
4
6
8

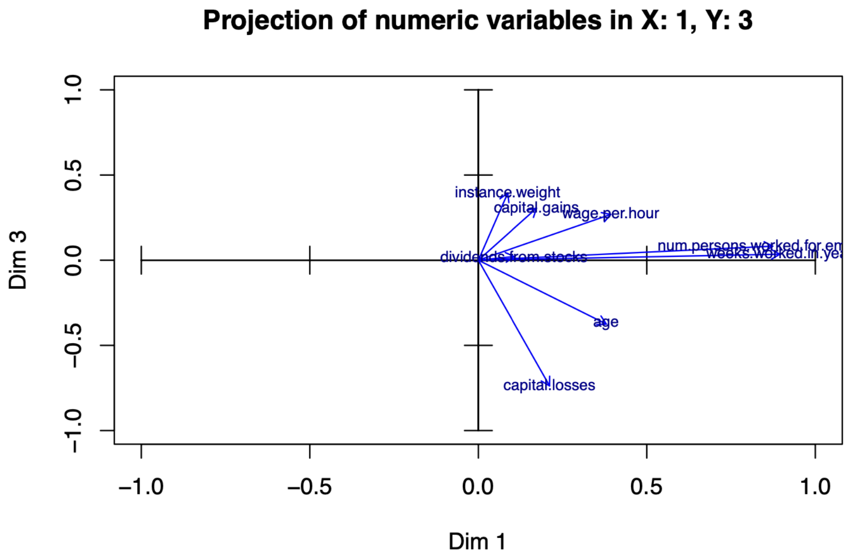

Individuals plot
PCA3 (12.67488460736 %)
60
40
20
0
−20
−40
−60
PCA2 (14.9647296009393 %)
−40
−30
−20
−10
0

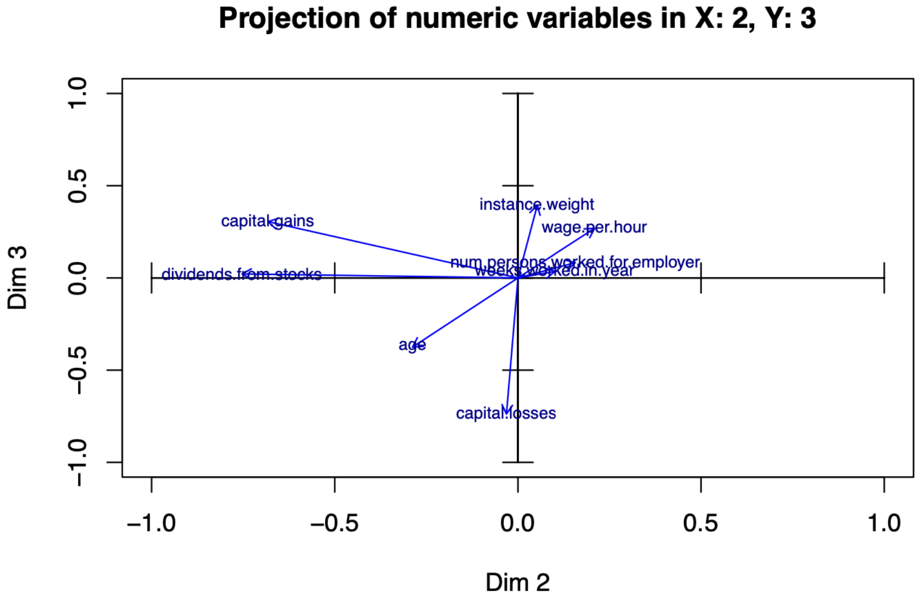
5.2.2 Common projection of numerical variables and modalities

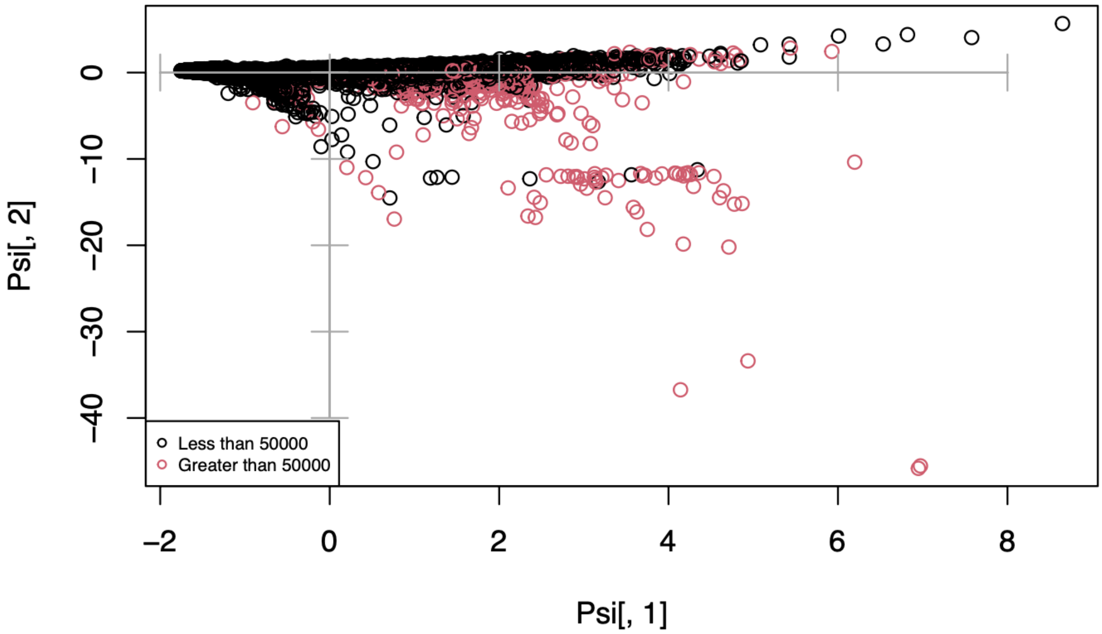
The common projection of numerical variables and modalities is a visualization of the variables in the two-dimensional space. Each variable is represented as an arrow pointing in the direction of the most important component, and the length of the arrow indicates the importance of the variable in that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

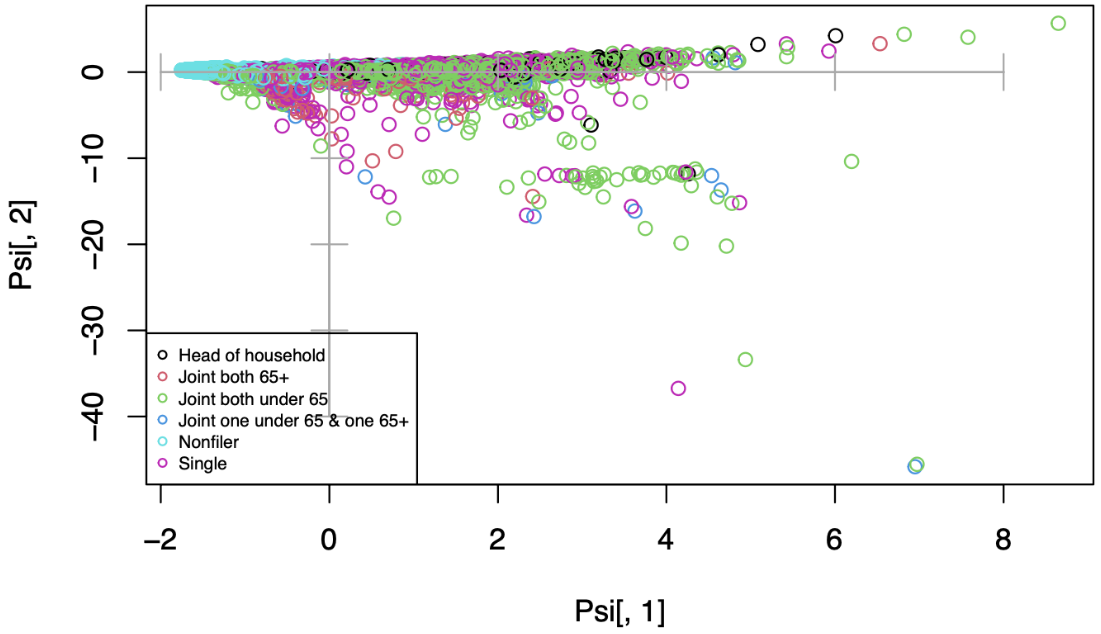


From this plot, we can see that the most important variables on the y-axis are dividends from stocks and capital gains. In terms of the x-axis, the most important are the weeks worked in a year and the number of persons worked for the employer.




Projection of numeric variables in X: 2, Y: 3
Dim 3
Dim 2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
instance.weight
capital.gains
wage.per.hour
num.persons.worked.for.employer
weeks.worked.in.year
dividends.from.stocks
age
capital.losses


Psi[, 2]
Psi[, 1]
0
−10
−20
−30
−40
−2
0
2
4
6
8
○ Less than 50000
○ Greater than 50000



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.