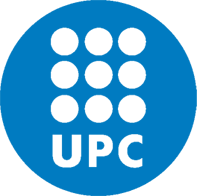
DELIVERABLE 1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ADEI Q2 24/25

****

****

Index

[**1. INTRODUCTION 2**](#_uilyebammyzp)

[**2. DATA SOURCE 2**](#_mmsmfargqidq)

[**3. DATA DESCRIPTION (METADATA) 2**](#_t7aemufuddxt)

[3. 1. DESCRIPTION OF THE VARIABLES 2](#_orwfwib51rj)

[**4. BASIC DESCRIPTIVE 6**](#_nzijvvi678ff)

[**5. PREPROCESSING 15**](#_wnbt205p59gg)

[5.1 VARIABLE SELECTION 15](#_zdmf76tryyb2)

[5.2 VARIABLE RENAMING 15](#_hniplpu8wywi)

[5.3 CATEGORY RENAMING 15](#_nahrmrvwqull)

[5.4 OUTLIERS 17](#_gn46ovcbfz77)

[5.5 MISSING VALUES 17](#_6ptse1ejzwb)

[**6. BASIC DESCRIPTIVE AFTER PREPROCESSING 18**](#_4w9oetui8jpu)

[**ANNEX: TAULA RESUM METADATA 26**](#_q08yyzv8c7ut)

# **1. INTRODUCTION**

Income classification plays a crucial role in socio-economic studies, helping to identify patterns and disparities in earnings. This dataset provides demographic and employment-related information, allowing for the analysis of factors influencing income levels. By exploring these variables, we can develop predictive models and gain insights into the relationships between education, occupation, work hours, and earnings.

# **2. DATA SOURCE**

The dataset used in this study comes from the **UCI Machine Learning Repository** and is available at the following link:  [Adult Census Income Dataset](https://www.kaggle.com/datasets/uciml/adult-census-income/data).

The primary objective is to predict whether an individual's income exceeds **$50K per year ("income" variable).**

# **3. DATA DESCRIPTION (METADATA)**

The selected dataset contains **48,842 instances** and **15 variables** that provide demographic and employment-related information about individuals.

## **3. 1. DESCRIPTION OF THE VARIABLES**

**AGE:**

* It is the number of years that a person had at that moment.
* Variable type: **integer.**

**WORKCLASS:**

* It is the name of the job, for example police, or teacher, in some rows we will see Private as a workclass, that means that this person is working in the private sector.
* Variable type: **categorical**

**FNLWGT:**

* Final weight assigned to the person, representing the estimated number of people with similar characteristics. Used for sample weighting but not typically relevant for predictive modeling.
* Variable type: **integer**

**EDUCATION:**

* The highest level of education achieved (e.g., Bachelors, HS-grad, Masters).
* Variable type: **categorical**

**EDUCATIONAL-NUM:**

* Numerical representation of the type of education that the person has been studying.
* Variable type: **integer**

#### \*\* Translation number - education

1 Preschool

2 1st-4th

3 5th-6th

4 7th-8th

5 9th

6 10th

7 11th

8 12th

9 HS-grad

10 Some-college

11 Assoc-voc

12 Assoc-acdm

13 Bachelor

14 Master

15 Prof-school

16 Doctorate

Since this dataset is from the US, we deem it interesting to give a further explanation of the categories that differ from our educational system.

**Assoc-voc:** Associate’s degree (vocational) a two-year college degree that focuses on practical job skills.

**Assoc-acdm:** Associate’s degree (academical) a two-year college degree, but it's more focused on general academic studies. It's often a stepping stone towards a four-year bachelor's degree.

**Prof-school:** Advanced education for specific professions, like being a doctor, lawyer, or dentist.

**MARITAL-STATUS:**

* Marital status of the individual (e.g., Married, Single, Divorced).
* Variable type: **categorical**

**OCCUPATION:**

* The type of occupation (e.g., Exec-managerial, Craft-repair, Sales).
* Variable type: **categorical**

**RELATIONSHIP:**

* Relationship status within a household (e.g., Husband, Wife, Own-child).
* Variable type: **categorical**

**RACE:**

* The race of the individual (e.g., White, Black, Asian-Pac-Islander).
* Variable type: **categorical**

**GENDER:**

* The gender of the individual (Male or Female).
* Variable type: **categorical**

**CAPITAL-GAIN:**

* Income from capital gains (e.g., profits from asset sales).
* Variable type: **integer**

**CAPITAL-LOSS:**

* Losses incurred from asset sales.
* Variable type: **integer**

**HOURS-PER-WEEK:**

* Number of hours worked per week.
* Variable type: **integer**

**NATIVE-COUNTRY:**

* Country of origin (e.g., United States, Mexico, Canada).
* Variable type: **categorical**

**INCOME (Target Variable):**

* Income category, indicating whether the individual earns **">50K" or "<=50K" per year.**
* Variable type: **categorical**

Finally, a summary of all the metadata has been made, located in the annex, you can go directly by clicking on this link : [ANNEX: TAULA RESUM METADATA](#_q08yyzv8c7ut)

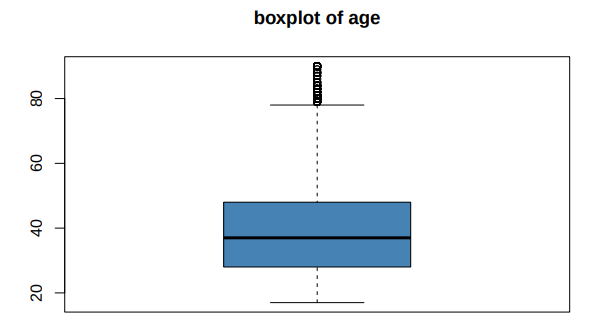
# 

# 4. BASIC DESCRIPTIVE

## 

* **AGE**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 17.0 |
| 1st Qu. | 28.0 |
| Median | 37.0 |
| Mean | 38.646 |
| 3rd Qu. | 48.0 |
| Max. | 90.0 |
| NA's | - |

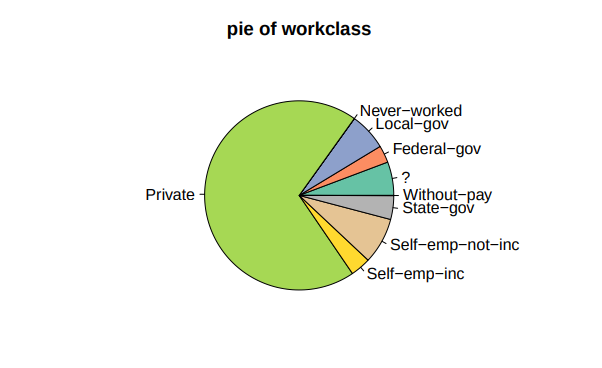
****

As we see in the table, the age distribution is slightly right-skewed, with a mean slightly higher than the median. This could indicate that while most individuals are younger, there are some older individuals pulling the average up.

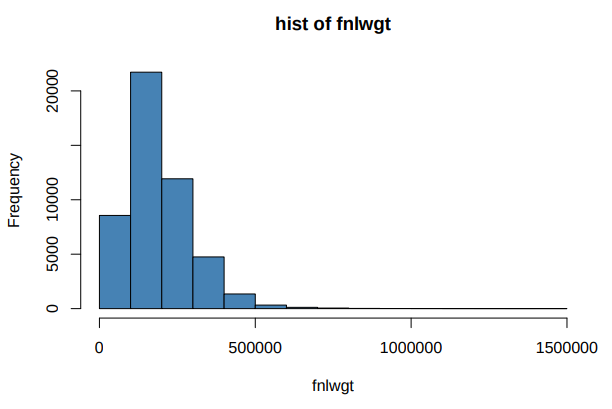
The histogram shows a concentration of individuals between 20 and 50 years old, with a peak around 30-40 years. This might be typical for a working-age population, where most people are in their prime working years. Although, the presence of individuals of 17 years old suggests that some people start working early, possibly in part-time or low-skill jobs. Also, the upper range (up to 90 years) indicates that some individuals continue working past the traditional retirement age.

From the boxplot we can confirm that we have a few outliers at the higher end of the age range, which represent older individuals who are still active in the workforce.

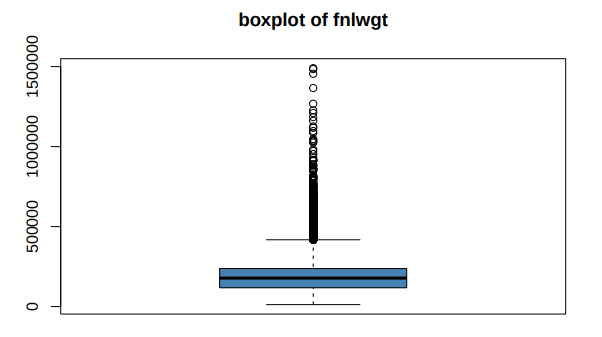
* **WORKCLASS**



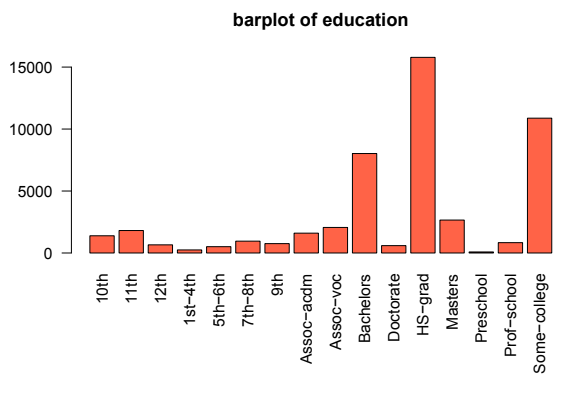
The barplot shows that the majority of individuals work in the private sector ("Private"), which might confirm the U.S. labor market structure, where the private sector is the largest employer. The pie chart visually reinforces the dominance of the private sector, with "Private" taking up the largest slice. Government and self-employed categories are smaller but still significant.

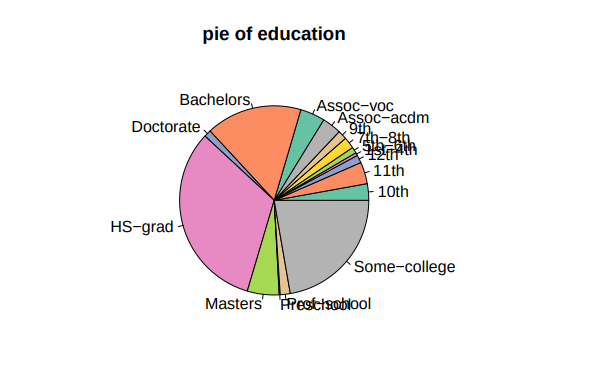
* **FNLWGT**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 12285.0 |
| 1st Qu. | 11755.0 |
| Median | 178144.0 |
| Mean | 189664.0 |
| 3rd Qu. | 237642.0 |
| Max. | 1490400.0 |
| NA's | - |



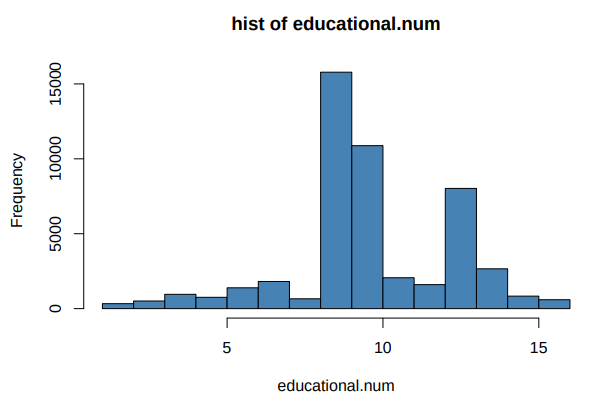
The histogram shows a highly skewed distribution, with most values concentrated at the lower end but with a long tail extending to very high values (up to 1,490,400). This reflects the nature of the final weight, which is used for population weighting. Also, the boxplot confirms the extreme skewness, with many outliers at the higher end of the scale. These outliers might represent individuals with very high final weights, which are used to adjust for underrepresented groups in the sample.

* **EDUCATION**

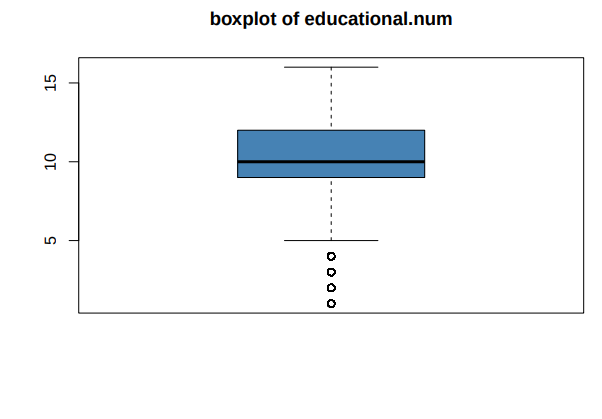


The barplot shows that the most common education levels are "HS-grad" (high school graduate) and "Some-college" (some college education). This reflects that many people complete high school but do not necessarily finish college. Also, the pie chart visually reinforces the dominance of high school graduates and those with some college education. Higher education levels like "Bachelors" and "Masters" are smaller slices but still important.

* **EDUCATIONAL-NUM**

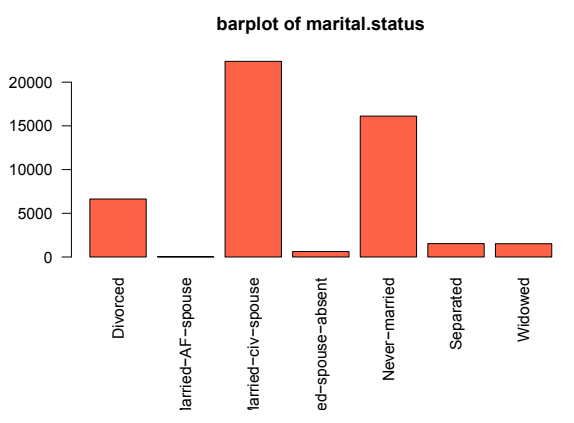
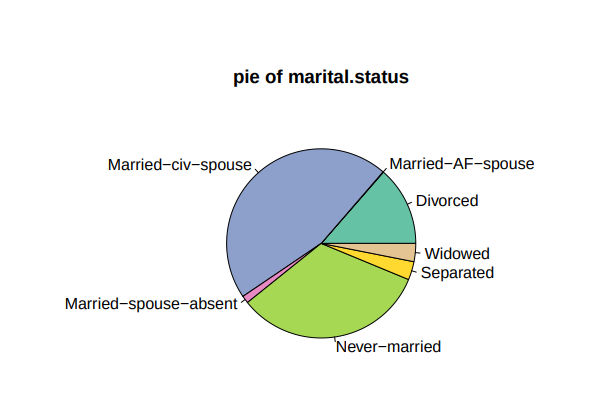


| **Statistical** | **Value** |
| --- | --- |
| Min. | 1.0 |
| 1st Qu. | 9.0 |
| Median | 10.0 |
| Mean | 10.1 |
| 3rd Qu. | 12.0 |
| Max. | 16.0 |
| NA's | - |

****

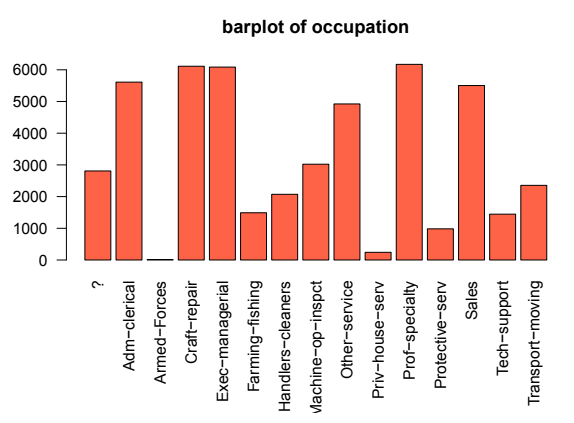
The histogram shows peaks at values corresponding to "HS-grad" (9) and "Some-college" (10), which aligns with the bar chart for the "EDUCATION" variable .Also, the boxplot shows a relatively symmetric distribution, with most values clustered around the median of 10, and there are no significant outliers.However, there is a slight skew toward higher education levels, with some individuals holding advanced degrees (up to 16, representing a doctorate)

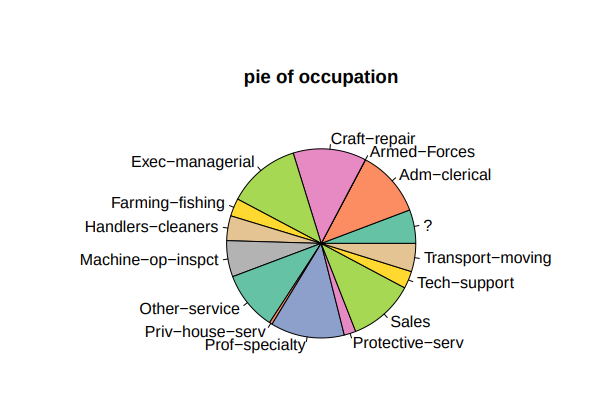
* **MARITAL-STATUS**



The barplot shows that the most common marital status is "Married-civ-spouse" (married, civilian spouse), followed by "Never-married" (single). This reflects the typical marital status distribution in the U.S., where marriage is common but a significant portion of the population remains single. Also, the pie chart visually reinforces the dominance of married individuals, with "Married-civ-spouse" taking up the largest slice, and Single individuals ("Never-married") are the next largest group.

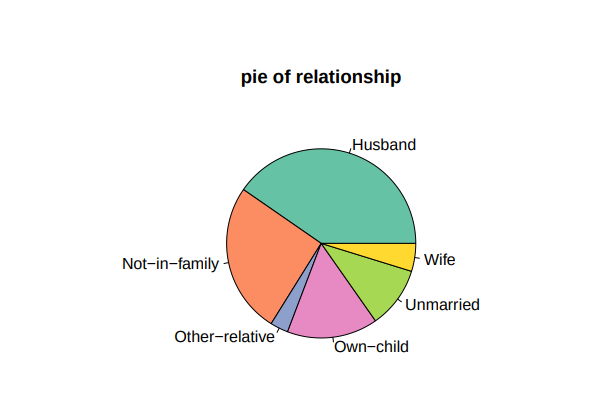
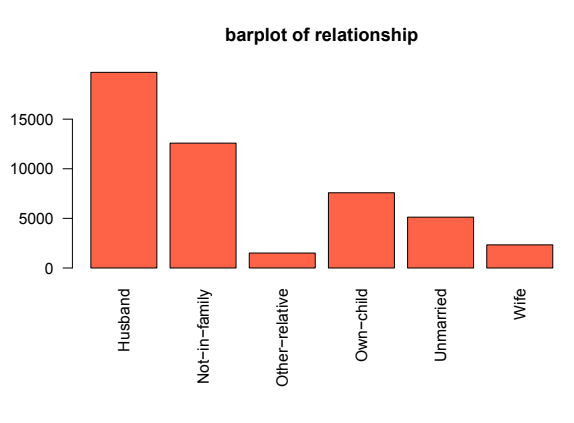
* **OCCUPATION**





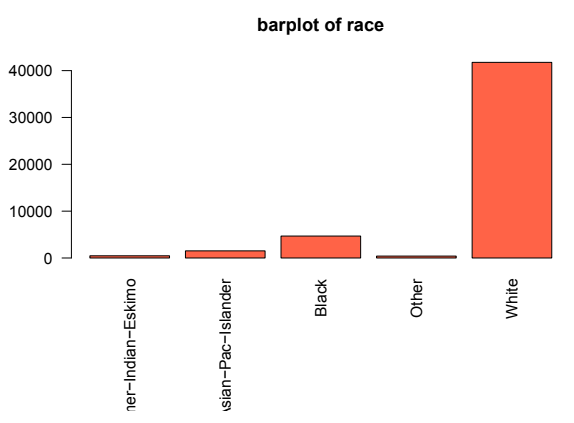
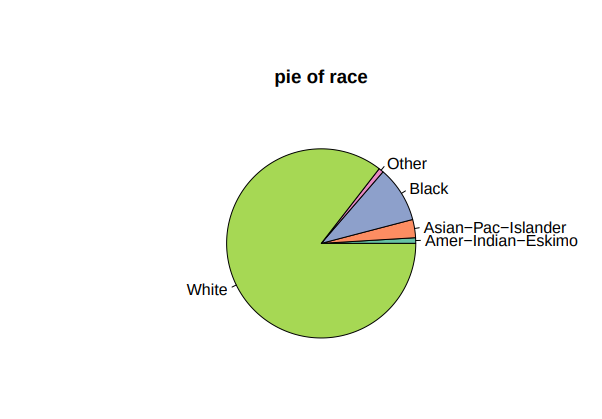
The barplot shows that the most common occupations are "Prof-specialty" (professional specialties) and "Craft-repair" (craft and repair), which reflects a diverse labor market, with a mix of skilled and unskilled jobs. Also, the pie chart visually reinforces the dominance of professional and craft-related occupations. Executive and managerial roles are smaller but still significant.

* **RELATIONSHIP**

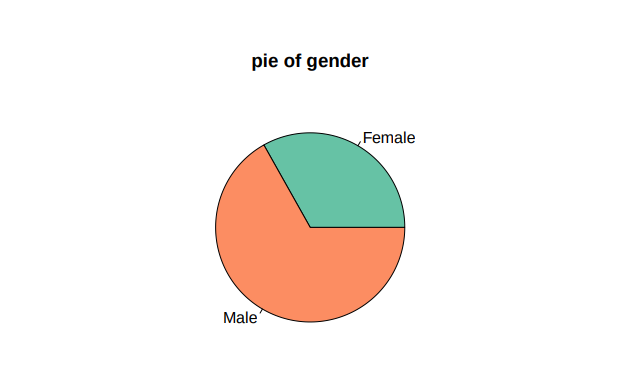


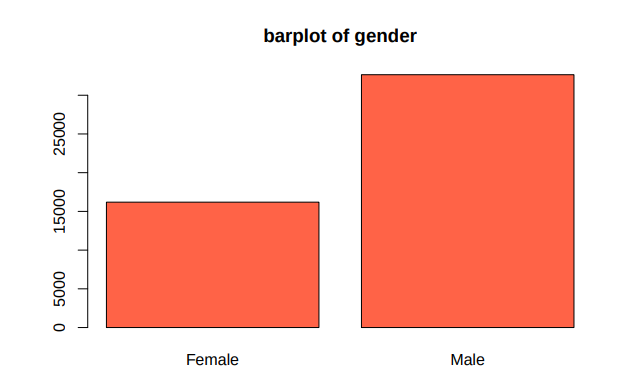
The barplot shows that the most common relationship status is "Husband," followed by "Not-in-family" (not in a family), which reflects that married couples are common in the USA. Also, the pie chart visually reinforces the dominance of "Husband" as the most common relationship status. "Not-in-family" is the next largest group.

* **RACE**

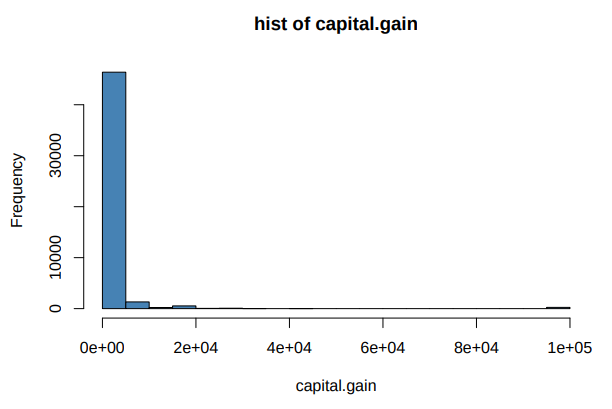


The barplot shows that the majority of individuals are White, followed by Black and Asian-Pac-Islander. This reflects the racial composition of the U.S. population in 1994. Also, as we can see, the pie chart visually reinforces the dominance of the White population, with smaller slices for Black and Asian-Pac-Islander individuals.

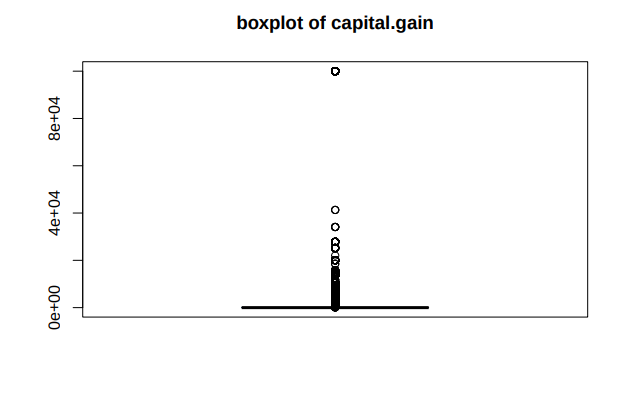
* **GENDER**



The bar chart shows that there are more males than females in the dataset. This reflects that men were more likely to be employed than women. Also, the pie chart visually reinforces the dominance of males in the dataset, with a smaller slice for females.

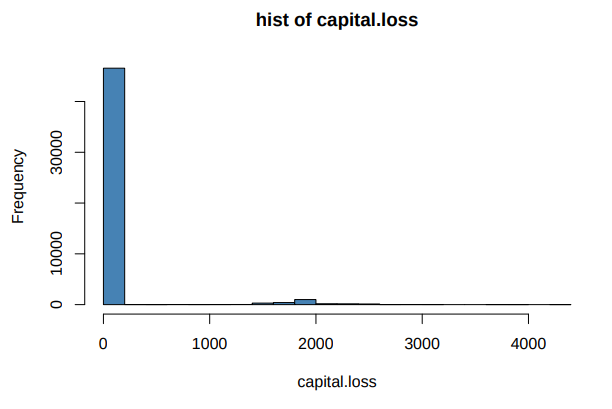
* **CAPITAL GAIN**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0.0 |
| 1st Qu. | 0.0 |
| Median | 0.0 |
| Mean | 1079.0 |
| 3rd Qu. | 0.0 |
| Max. | 99999 |
| NA's | - |



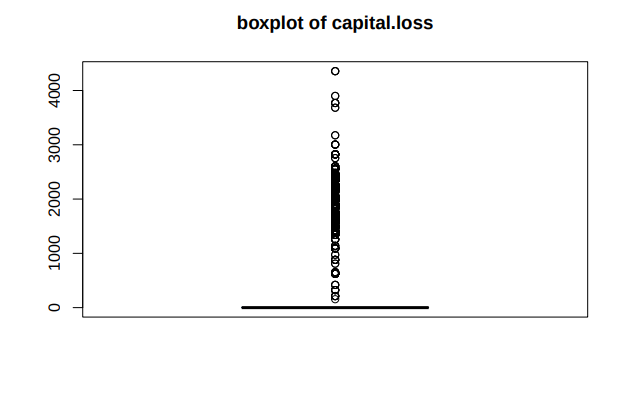
The histogram shows that most people have no capital gains (value is 0) with a long tail extending to very high values. However, those who do have capital gains tend to have higher incomes, as capital gains are often associated with investments or assets that generate additional income.

This suggests that capital gains could be an indicator of a higher socioeconomic status and, therefore, a higher salary. Also, as we can see in the table the median and first and third quartiles are 0, confirming that most people have no capital gains. However, there are some extreme values (up to 99,999), indicating that some individuals have significant gains.



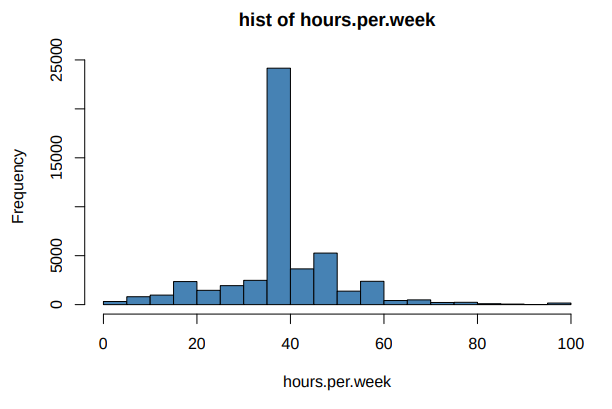
* **CAPITAL LOSS**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 0.0 |
| 1st Qu. | 0.0 |
| Median | 0.0 |
| Mean | 87.5 |
| 3rd Qu. | 0.0 |
| Max. | 4356.0 |
| NA's | - |

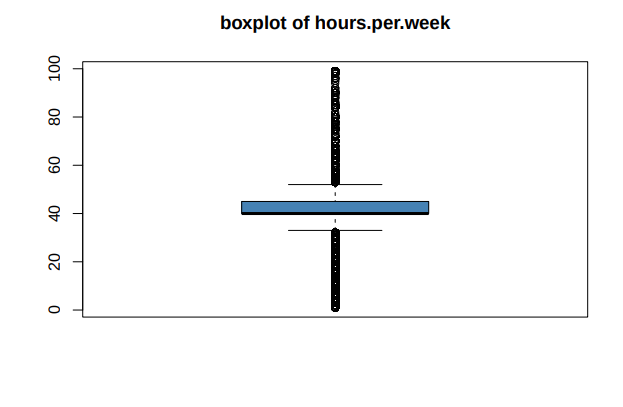


Similar to capital gains, the histogram shows that most individuals have zero capital losses, with a long tail extending to higher values. This indicates that capital losses are rare but can be significant for those who incur them.

However, those who do have capital losses might be in a less stable financial situation, which could correlate with lower incomes. Capital losses can occur when individuals sell assets (e.g., properties or stocks) at a price lower than what they paid for them. Also, as we can see in the table, the median and first and third quartiles are 0, confirming that most people have no capital losses. However, there are some extreme values (up to 4,356), indicating that some individuals have significant losses.

* **HOURS PER WEEK**

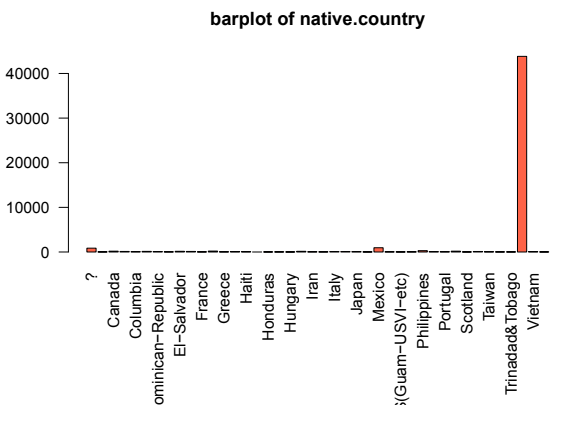
| **Statistical** | **Value** |
| --- | --- |
| Min. | 1.0 |
| 1st Qu. | 40.0 |
| Median | 40.0 |
| Mean | 40.42 |
| 3rd Qu. | 45.0 |
| Max. | 99.0 |
| NA's | - |

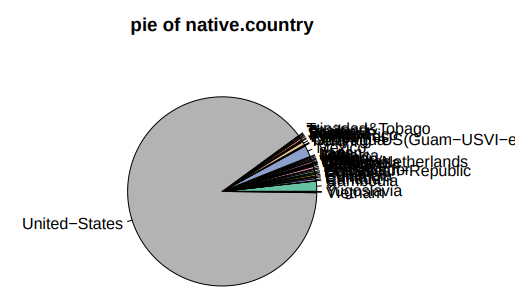


The histogram shows a peak at 40 hours, with a few individuals working significantly more or fewer hours, which reflects the standard for full-time employment.

There are some extreme values (up to 99 hours), indicating that some individuals work very long hours. This could reflect individuals with multiple jobs or those working in sectors that require long hours, such as medicine or law. Also,as we can see in the table the median is 40 hours, confirming that most people work full-time. The mean is 40.42 hours, suggesting a slight right skew in the distribution due to some individuals working many hours.

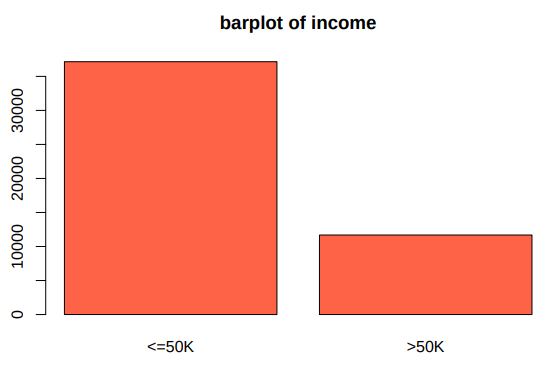
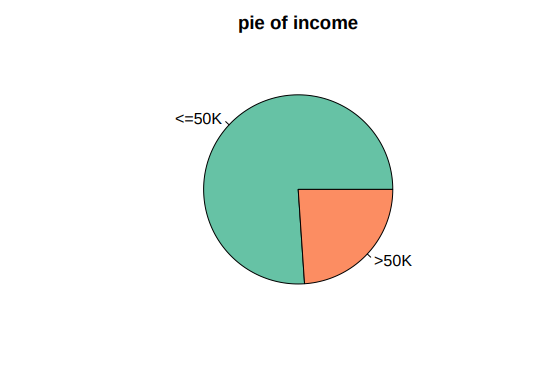
* **NATIVE COUNTRY**





The barplot shows that the majority of individuals are from the United States. Other countries are represented but in much smaller numbers.. In the dataset, immigrants tended to have lower incomes than native-born individuals, especially if they lacked higher education or English proficiency. Individuals from other countries, such as Mexico, Canada, and various Asian and European countries, are also observed, though in smaller proportions.

* **INCOME**



The barplot shows that most people have an income of less than or equal to 50K ("<=50K"), while a minority have an income greater than 50K (">50K"). This reflects that most people earned modest incomes. So, as we can see individuals with higher incomes tend to have higher education levels, better-paying occupations, and more work experience. Also, the pie chart visually reinforces the dominance of lower-income individuals, with a smaller slice for those earning more than 50K.

# **5. PREPROCESSING**

## **5.1 VARIABLE SELECTION**

In our dataset we have got a total of 15 variables. Nevertheless, there are two of them that share the same information, education and educational.num. In order not to have it repeated, we only keep the numeric one (educational.num). The rest remains unchanged.

## **5.2 VARIABLE RENAMING**

We decided to rename some variables to make the data more readable. The renames are:

* educational-num → edu\_num
* marital-status → marital
* gender → sex
* capital-gain → cap\_gain
* capital-loss → cap\_loss
* hours-per-week → hours\_week
* native\_country → native\_country

## **5.3 CATEGORY RENAMING**

A similar process is done in data categories:

**Variable:** Workclass

* Private → Priv
* Self-emp-not-inc →SelfN
* Self-emp-inc → SelfI
* Federal-gov → Fed
* Local-gov →Loc
* State-gov → State
* Without-pay → NoPay
* Never-worked → NoPay

*SelfI means incorporated (in a company) and selfn(no inc) means self-employed for instance as a freelancer.*

We decided to join the categories Without-pay and Never-worked given there are few registers and the objective of the dataset is to predict whether the individual earns more or less than 50k$ a year and both groups earn 0.

**Variable:** Marital

* Never-married → NevMarr
* Married-civ-spouse → Married
* Married-AF-spouse → Married
* Separated → Sep
* Divorced → Div
* Widowed → Widow

Married-civ-spouse and Married-AF-spouse are merged as both represent the same legal married status and distinguishing between civil and military spouses does not provide significant analytical value. This simplifies the dataset without losing relevant information.

**Variable:** Occupation

* Exec-managerial → ExecMan
* Prof-specialty → Prof
* Adm-clerical → AdminCler
* Craft-repair →CraftRep,
* Transport-moving → Trans
* Handlers-cleaners → HandlCl
* Machine-op-inspct →MachOp
* Tech-support → Tech
* Protective-serv → ProtServ
* Armed-Forces → Army
* Farming-fishing → FarmFish
* Priv-house-serv → House
* Other-service → Other

These renaimings provides clarity, consistency and efficiency in data analysis. The old names are too long and contain hyphens, making them harder to process and visualize.

**Variable:** Native\_country

* United-States → USA
* Different countries which are not United-States → Other

As there very few registers from people born in countries other than the USA (>1000), we decided to join all these countries in the category “other”.

## **5.4 OUTLIERS**

To check for outliers we used two different methods. We detected outliers using the Interquartile Range (IQR) method and boxplots:

**IQR Method:**

We calculated the 25th percentile (Q1) and 75th percentile (Q3) for each numerical variable. The IQR was computed as Q3 - Q1, and values were flagged as outliers if they fell below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR.

**Boxplots:**

We visualized outliers using boxplots, where extreme values appeared as individual points outside the whiskers. This helped identify variables with potential anomalies.

This approach allowed us to identify anomalies; however, since these values represent extreme yet plausible cases within the population, we chose to leave them unchanged to maintain the natural variability in the data.

## **5.5 MISSING VALUES**

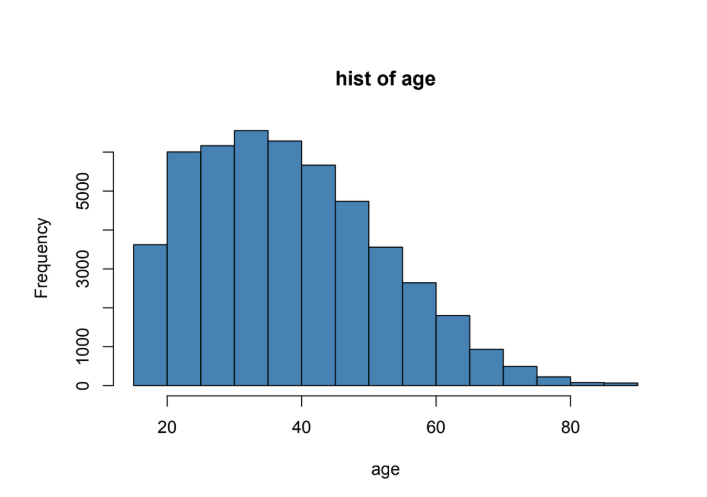
To handle missing values, we have chosen to fill them with the most frequent value. Since these missing values occur randomly, their absence does not convey meaningful information and is not due to the value being inherently inapplicable—such as recording "number of pregnancies" for people whose assigned gender at birth is male.

# 

# 6. BASIC DESCRIPTIVE AFTER PREPROCESSING

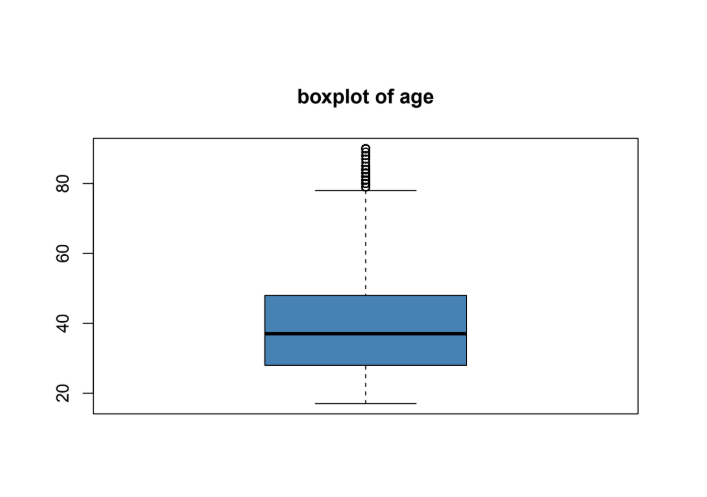
After the preprocessing process, the most significant changes are found in the variables Workclass, Marital, and Native\_country. In these variables, some categories have been renamed, merged, and even reduced.

In other variables, such as Age, no changes have been made, but it is important to highlight that, despite observing outliers, it was decided not to address them as they are plausible and do not negatively affect the analysis since they represent real cases and have meaning. No impossible values or errors were found.

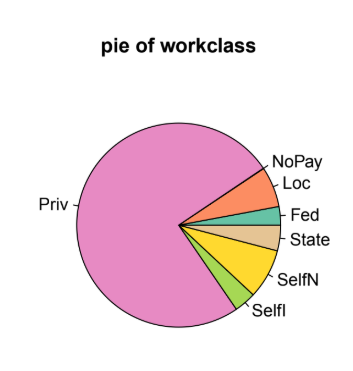
Finally, when replacing unknown values with the most frequent values, no significant changes were observed.

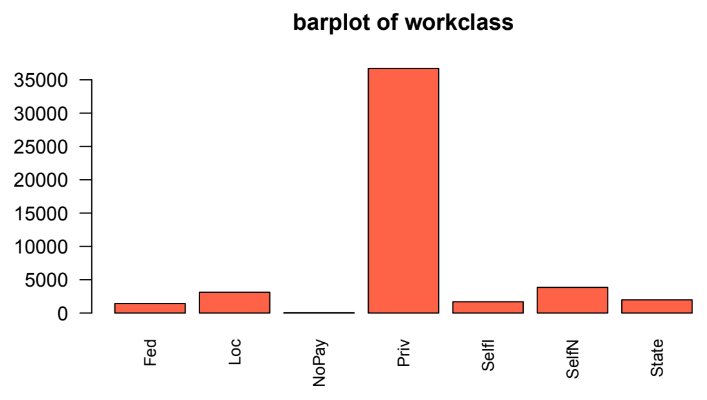
* **AGE**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 17.0 |
| 1st Qu. | 28.0 |
| Median | 37.0 |
| Mean | 38.64 |
| 3rd Qu. | 48.0 |
| Max. | 90.0 |
| NA's | - |

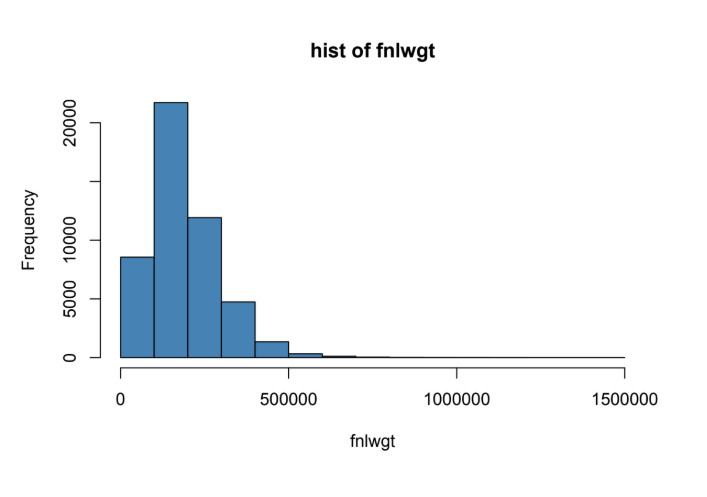


In the variable Age, no changes were made after preprocessing. Statistical values such as the minimum, maximum, quartiles, mean, and median remain identical. This means that histograms and boxplots do not show any alterations, maintaining the same distribution with a slight right skew and the presence of outliers at the upper extreme, corresponding to individuals who continue working at advanced ages.

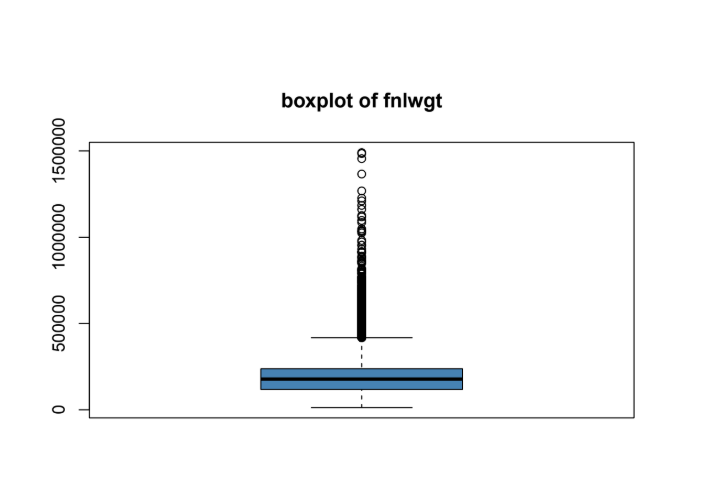
* **WORKCLASS**

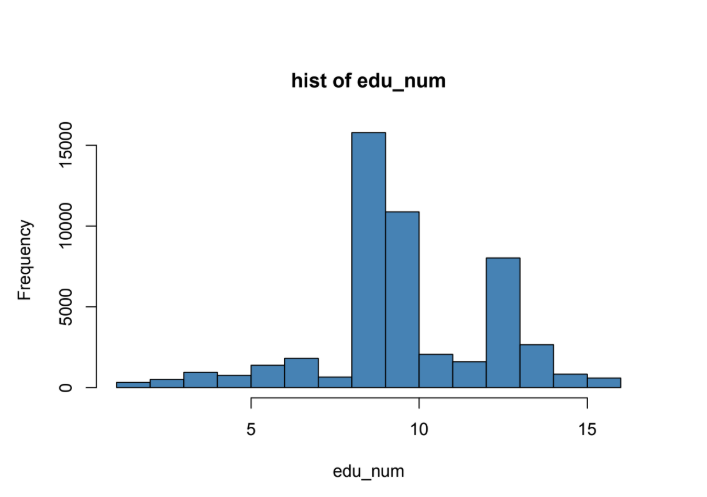


In the variable Workclass, modifications were made to the category names to make the classification clearer and more manageable. Some labels were shortened, such as Private to Priv, Self-emp-not-inc to SelfN, and Self-emp-inc to SelfI. Additionally, the categories Without-pay and Never-worked were grouped into a new category called NoPay, as both represent unemployed individuals. Graphically, these changes are reflected in a reduction in the number of categories in bar plots and pie charts, making data interpretation easier.

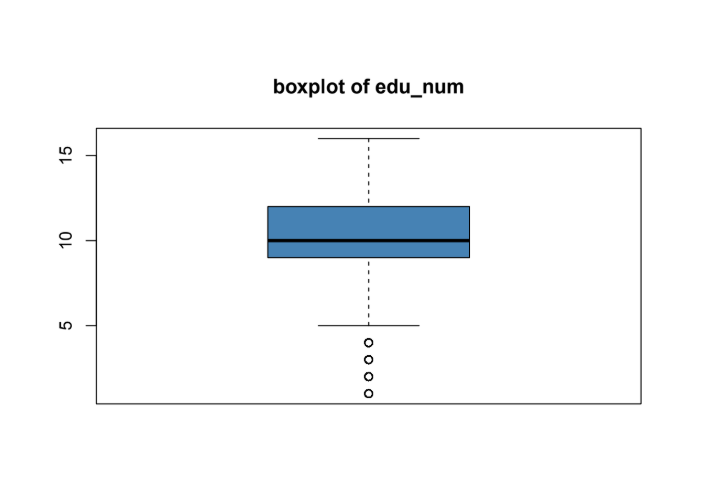
* **FNLWGT**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 12285.0 |
| 1st Qu. | 117550.0 |
| Median | 178144.0 |
| Mean | 189664.0 |
| 3rd Qu. | 237642.0 |
| Max. | 1490400.0 |
| NA's | - |

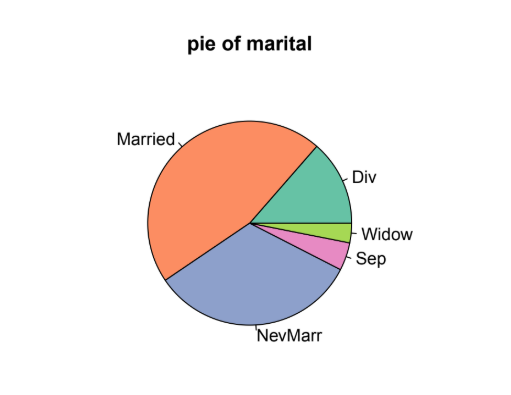
In the variable FNLWGT, which represents the final weight assigned to each person in the sample, no modifications were made during preprocessing. Descriptive statistics such as the mean, median, and extreme values remain the same, maintaining the highly skewed distribution with very high values at the upper extreme. This implies that histograms and boxplots for this variable show no differences after preprocessing.

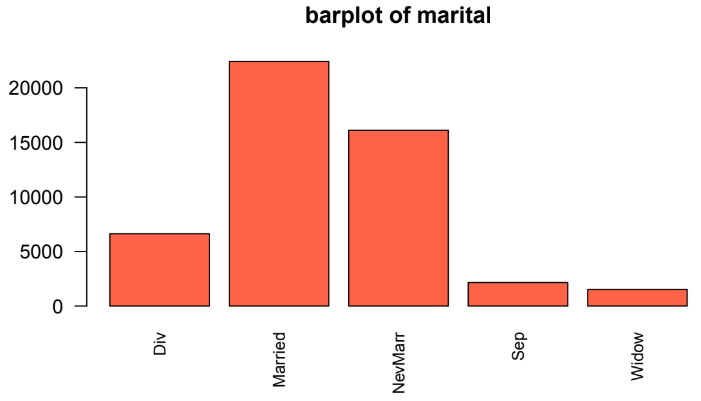
* **EDU-NUM**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 1.00 |
| 1st Qu. | 9.00 |
| Median | 10.00 |
| Mean | 10.08 |
| 3rd Qu. | 12.00 |
| Max. | 16.00 |
| NA's | - |



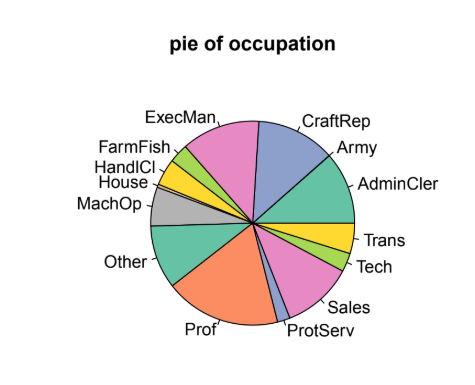
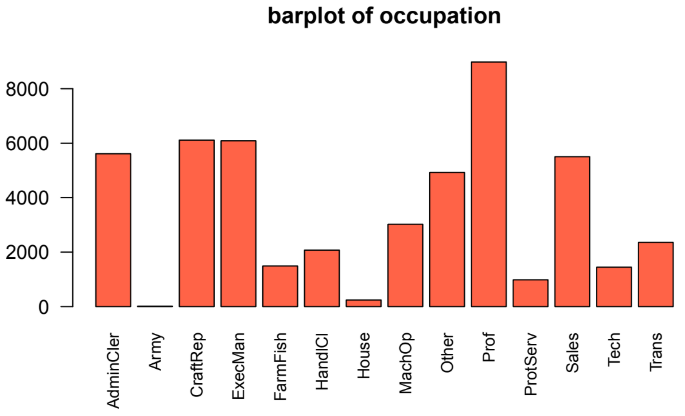
In the variable Education-num, renamed as Edu-num, statistical values remained almost identical after preprocessing. The median remains 10, and the mean changed slightly from 10.1 to 10.08, a minimal variation that does not significantly affect the distribution. Graphically, histograms and boxplots retain the same shape, with the highest concentration of values corresponding to HS-grad and Some-college education levels.

* **MARITAL**

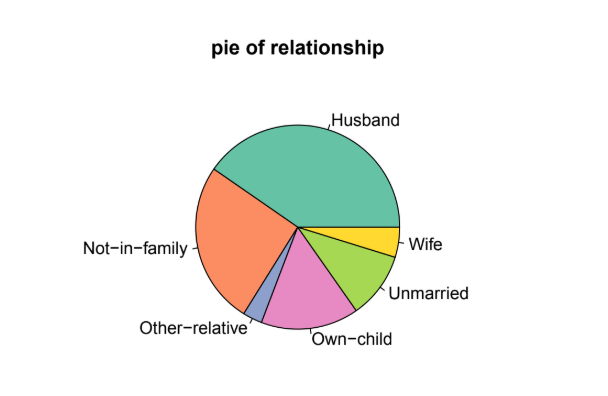


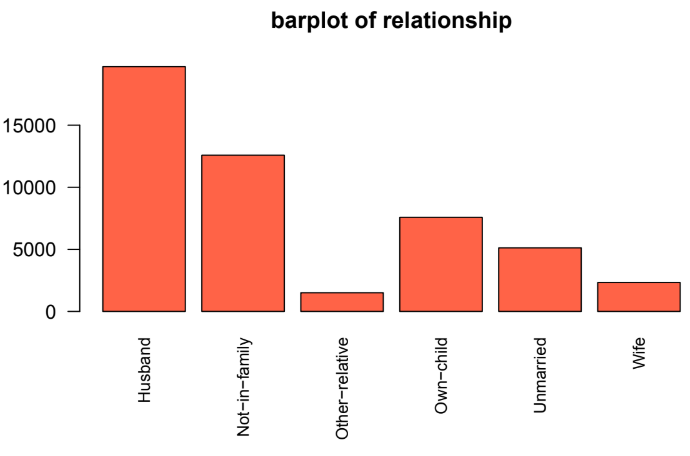
In the variable Marital, significant changes were made by merging and renaming some categories. The labels Married-civ-spouse and Married-AF-spouse were combined into a single category called Married, as both represent the married civil status, and distinguishing between them does not provide analytical value. Other labels were also shortened and simplified, such as Never-married to NevMarr, Separated to Sep, Divorced to Div, and Widowed to Widow. These modifications are reflected in bar plots and pie charts, where the number of categories has been reduced, making the data easier to read and interpret.

* **OCCUPATION**

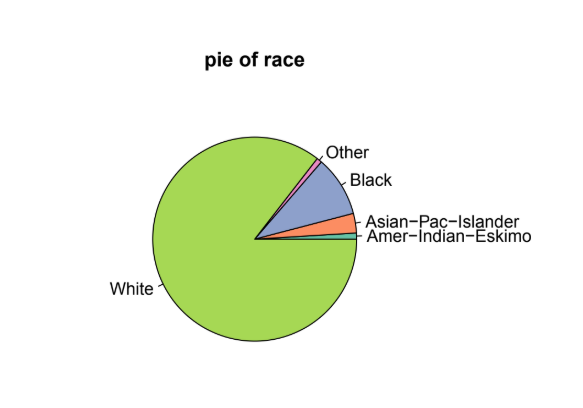


In the variable Occupation, changes were made to category names to make them shorter and more understandable. Some new labels include ExecMan instead of Exec-managerial, Prof instead of Prof-specialty, and AdminCler instead of Adm-clerical, among others. Additionally, similar occupations were grouped to simplify categorization. These changes are visible in bar plots and pie charts, where the number of categories has been reduced, improving the visual clarity of the occupation distribution.

* **RELATIONSHIP**

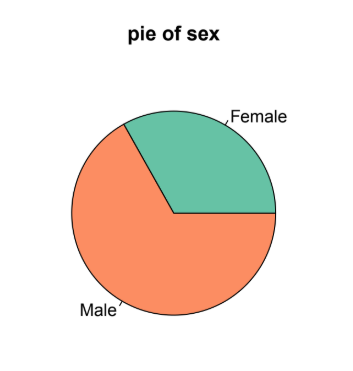


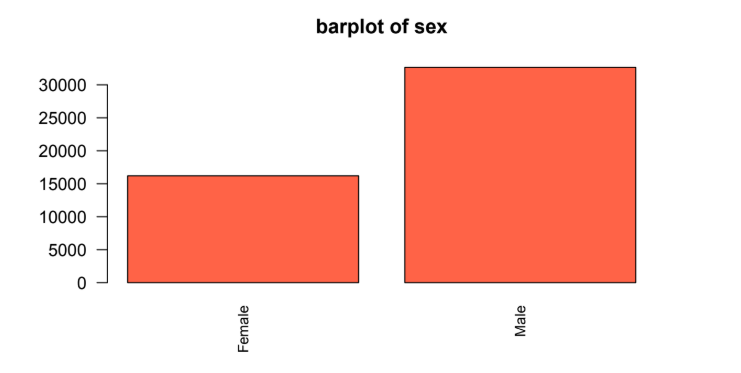
In the variable Relationship, no changes were made during preprocessing, so descriptive statistics and data distribution remain the same. Bar plots and pie charts also show no differences, as the variable retains the same categories and proportions.

* **RACE**

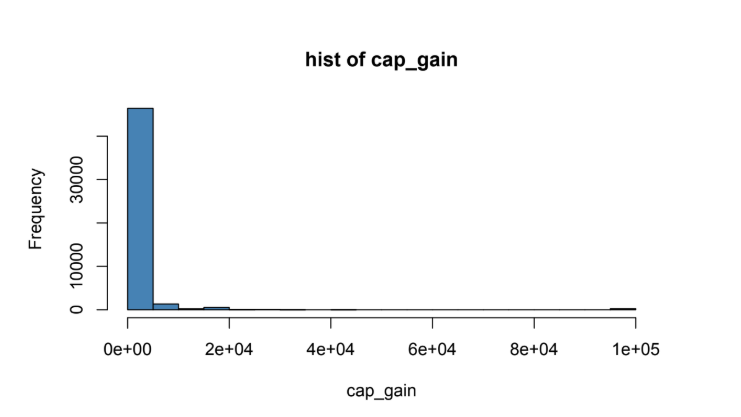
# 

In the variable Race, no modifications were made during preprocessing either. The data distribution still shows that the majority of individuals are White, followed by Black and Asian-Pac-Islander, without any changes in frequencies. Graphically, bar plots and pie charts maintain the same structure and proportions as in the original dataset.

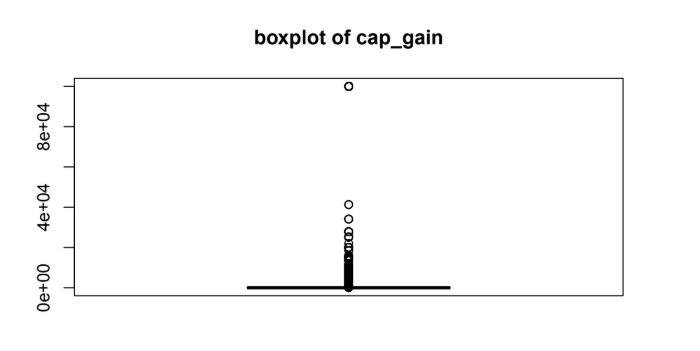
* **SEX**



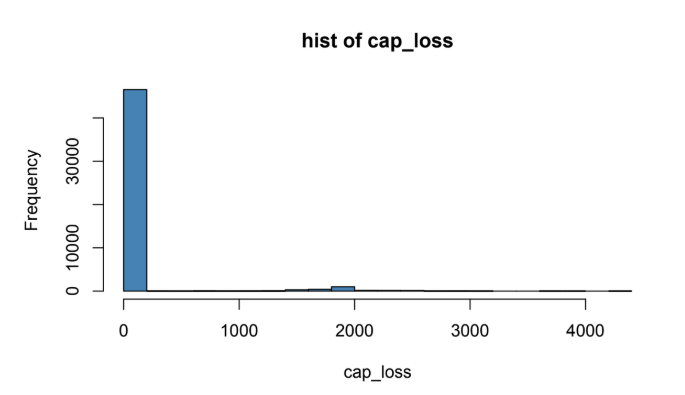
In the variable Sex, previously called Gender, only a name change was made without modifying the data itself. This means that histograms, boxplots, and other distribution charts remain identical, maintaining the predominance of the Male category over Female.

* **CAP-GAIN**

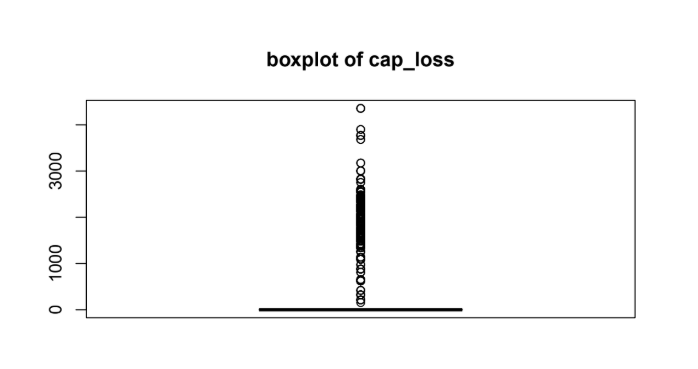
| **Statistical** | **Value** |
| --- | --- |
| Min. | 0 |
| 1st Qu. | 0 |
| Median | 0 |
| Mean | 1079 |
| 3rd Qu. | 0 |
| Max. | 99999 |
| NA's | - |



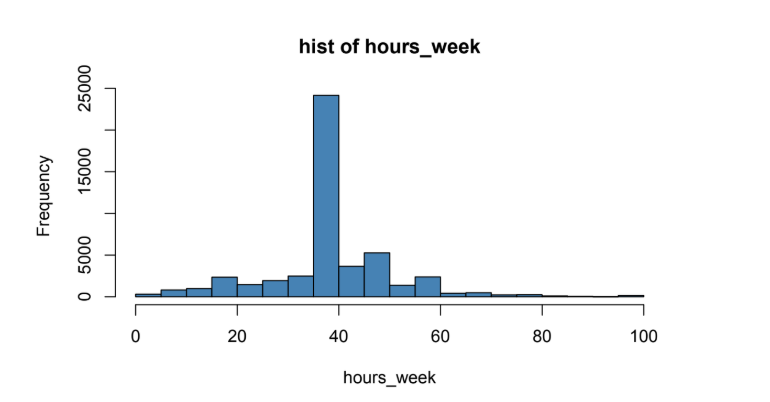
In the variable Capital-gain, no modifications were made during preprocessing. The distribution remains highly skewed, with most individuals having a value of zero and only a few with extremely high values. The histograms and boxplots continue to show a long tail toward high values, confirming that capital gains are concentrated among a small portion of the population.

* **CAP-LOSS**

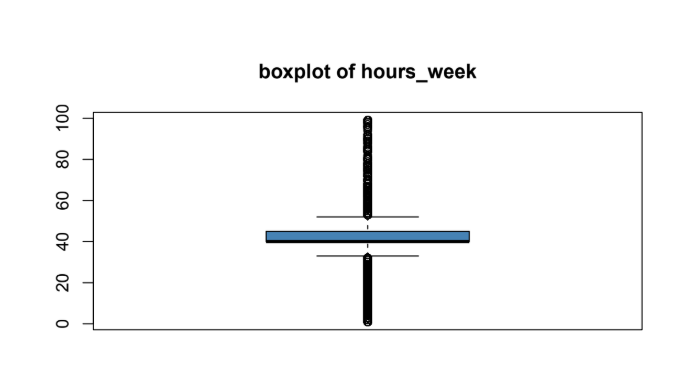
| **Statistical** | **Value** |
| --- | --- |
| Min. | 0.0 |
| 1st Qu. | 0.0 |
| Median | 0.0 |
| Mean | 87.5 |
| 3rd Qu. | 0.0 |
| Max. | 4356.0 |
| NA's | - |



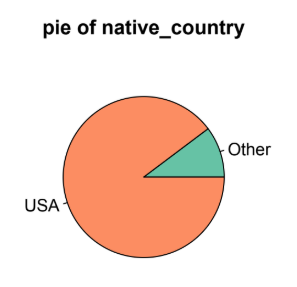
In the variable Capital-loss, no changes were applied during preprocessing. The distribution is similar to Capital-gain, with most values at zero and a few extreme cases with high losses. The histograms and boxplots maintain the same structure, with a right-skewed distribution and a long tail toward higher values.

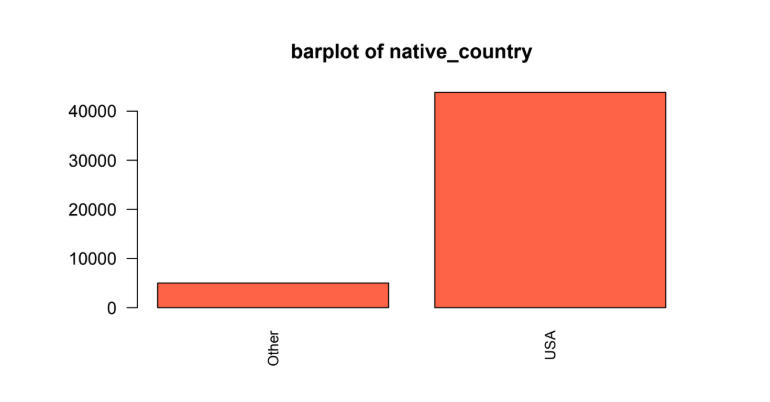
* **HOURS-WEEK**

| **Statistical** | **Value** |
| --- | --- |
| Min. | 1.00 |
| 1st Qu. | 40.00 |
| Median | 40.00 |
| Mean | 40.42 |
| 3rd Qu. | 45.00 |
| Max. | 99.00 |
| NA's | - |

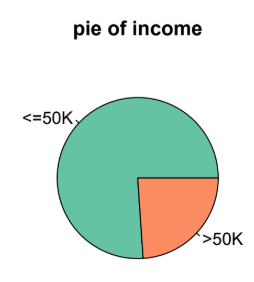


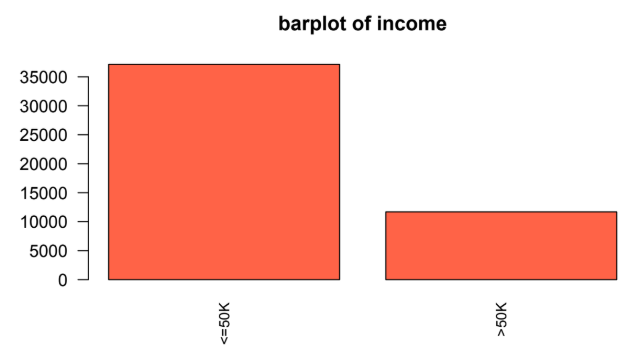
In the variable Hours per week, no changes were applied during preprocessing. The mean remains 40.42 and the median 40, indicating that most people work 40 hours per week, with some individuals significantly exceeding this value. In the graphs, the histogram and boxplot maintain the same distribution, with a peak at 40 hours and some outliers at the upper extreme corresponding to individuals who work very long hours.

* **NATIVE-COUNTRY**



In the variable Native Country, a major simplification was made by grouping all categories except the United States into a single category called Other. Before preprocessing, the variable had multiple categories with few records, making analysis difficult. After preprocessing, the bar plots and pie charts now display only two categories, USA and Other, making the variables easier to interpret.

* **INCOME**



In the variable Income, which is the target variable, no modifications were made during preprocessing. The distribution still shows that most individuals have an income of less than or equal to 50K, while only a minority exceed that threshold. The bar plots and pie charts reflect this strong asymmetry without changes from the original version.

# ANNEX: TAULA RESUM METADATA

| **VARIABLE NAME** | **SHORT VARIABLE** | **MEANING** | **TYPE** | **MIN** | **MAX** | **UNIT** | **MODALITY** | **SHORT MODALITY** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AGE | age | Number of years a person had at that moment | Integer | 17 | 90 | Years | - | - |
| WORKCLASS | workclass | Sector or type of work.In some rows we will see Private as a workclass, that means that this person is working in the private sector. | Categorical | - | - | - | - Private  Self-emp-not-inc  -Self-emp-inc  -Federal-gov  -Local-gov  -State-gov  -Without-pay -Never-worked | -Priv  -SelfN  -SelfI  -Fed  -Loc  -State  -NoPay  -NoPay |
| FNLWGT | fnlwgt | Final weight assigned to the person, representing the estimated number of people with similar characteristics. | Integer | 12285 | 1490400 | - | - | - |
| EDUCATION | education | The highest level of education achieved | Categorical | - | - | - | [**\*\***](#_j7pnxjmgj2x5) | - |
| EDUCATIONAL-NUM | edu\_num | Numerical representation of education level. | Integer | 1 | 16 | - | [**\*\***](#_j7pnxjmgj2x5) | - |
| MARITAL-STATUS | marital | Marital status of the individual. | Categorical | - | - | - | -Never-married  -Married-civ-spouse  -Married-AF-spouse  -Separated  -Divorced  -Widowed | -NevMarr  -Married  -Married  -Sep  -Div  -Widow |
| OCCUPATION | occupation | Type of occupation or profession. | Categorical | - | - | - | -Exec-managerial  -Prof-specialty  -Adm-clerical -Craft-repair  -Transport-moving -Handlers-cleaners  -Machine-op-inspct  -Tech-support -Protective-serv  -Armed-Forces  -Farming-fishing  -Priv-house-serv  -Other-service | -ExecMan  -Prof  -AdminCler  -CraftRep,  -Trans  -HandlCl  -MachOp  -Tech  -ProtServ  -Army  -FarmFish  -House  -Other |
| RELATIONSHIP | relationship | Relationship status within a household. | Categorical | - | - | - | -Husband  -Wife  -Own-child  -Not-in-family  -Other-relative  -Unmarried | - |
| RACE | race | Race of the individual. | Categorical | - | - | - | White, Black, Asian-Pac-Islander | - |
| GENDER (SEX) | sex | Sex of the individual. | Categorical | - | - | - | Male, Female | - |
| CAP-GAIN | cap\_gain | Income from capital gains (e.g., asset sales). | Integer | 0 | 99999 | USD | - | - |
| CAP-LOSS | cap\_loss | Losses incurred from asset sales. | Integer | 0 | 4356 | USD | - | - |
| HOURS-WEEK | hours\_week | Number of hours worked per week. | Integer | 1 | 99 | Hours | - | - |
| NATIVE-COUNTRY | native\_country | Country of origin. | Categorical | - | - | - | -United States  -Other | -USA  -Other |
| INCOME  (Target Variable) | income | Income category (>50K or <=50K per year). | Categorical | - | - | USD | >50K,  <=50K | - |