Presentation and Visualization

Census Income Analysis Dashboard: Visualization of the Adult Dataset

Ilaria Curzi, Georgia Zavou, Alana Zoloeva, Theodoros Lambrou

For this project, we have chosen the "Adults" dataset from the U.S. Census ("Census Income"), which focuses on predicting whether an individual's income exceeds 50,000 USD per year. This dataset, sourced from the official United States government website, was extracted by Barry Becker in 1994. Although certain elements may seem outdated, such as the inclusion of Yugoslavia, this dataset offers valuable insights for our analysis and dashboard creation, covering demographic, educational, and occupational aspects relevant to income levels.

Part 1: Exploratory Data Analysis (EDA): Data Overview and Relationship Exploration

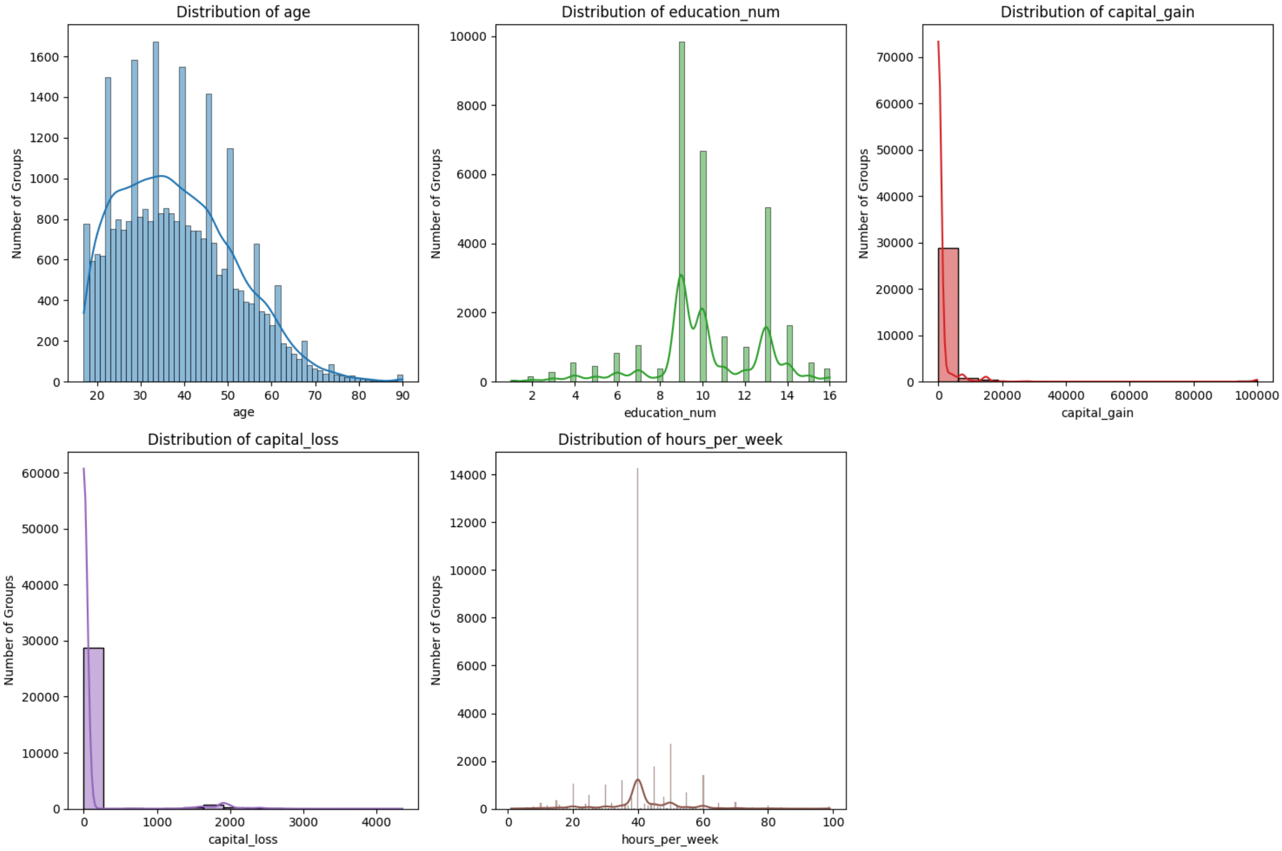
We begin by describing the dataset and exploring relationships within the data. The dataset consists of two main data types: categorical and numerical. Categorical features include "workclass," "education," "marital_status," "occupation," "relationship," "race," "sex," and "native_country," while numerical features encompass "age," "fnlwgt" (final weight), "education_num," "capital_gain," "capital_loss," and "hours_per_week." Overall, the dataset contains 14 features, from which we intend to select 5-6 key features for our dashboard visualizations.

Though the dataset lacks explicit temporal data and is not structured as a time series, it does contain geographical information. For instance, the “native_country” feature allows for potential geographic analysis based on the origin of individuals, revealing any geographic trends within the data.

Regarding data characteristics, the ranges, units, and precision of the numerical features vary according to each specific feature. Additionally, this dataset represents a fixed point in time, capturing a single snapshot rather than dynamic, updatable data, and can thus be treated as a static dataset.

The following series of histograms present the distribution of selected numerical features in the dataset, including age, education_num, capital_gain, capital_loss, and hours_per_week. These distributions help in understanding how different numerical features are spread and their relationship to the overall dataset.

Distribution Plots



We now describe the most important columns.

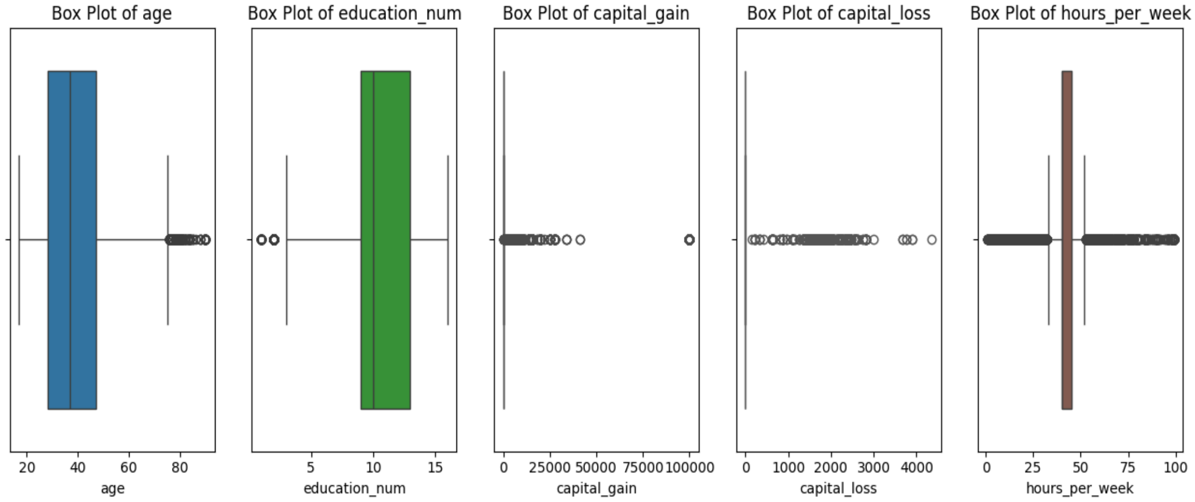
Age
The age distribution is slightly right-skewed, with the most common age group being in the mid-30s. This suggests that most people in the sample are in their 30s. The overall spread of ages is smooth, and the rightward skew is clearly visible in the data.

Hours per Week
The distribution of weekly working hours closely follows a normal pattern, with the peak around 40 hours per week, which is typical for full-time employment. This suggests that most people in the sample work around 40 hours a week.

Education (Years of Schooling)
The distribution of years of schooling is not normal, showing two peaks: one at 9 years and another at 13 years. This likely represents the completion of primary school and high school/university. The most common years of schooling are 9, 10, and 13 years.

Outliers

In the same way, following a specific focus on the "age" and "hours_per_week" features, we generated boxplots to identify and highlight potential outliers.



The boxplot for age reveals several outliers - a few individuals aged 75 and older. These outliers indicate the presence of older individuals who fall outside of the typical age range of the majority of the dataset, suggesting a small group of participants who are significantly older than the rest.

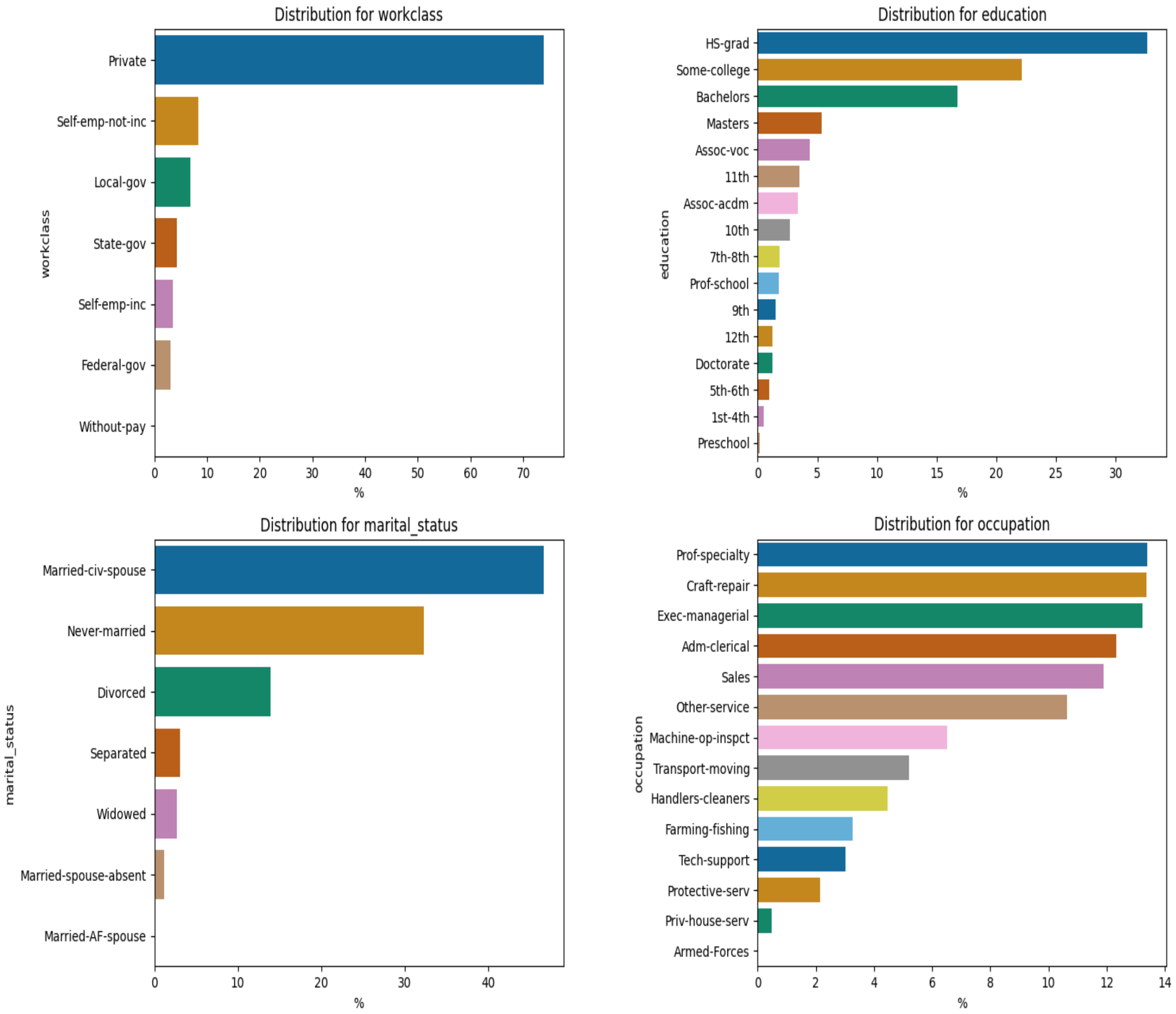
- Years of schooling (education_num): Shows a few outliers below 5 years of education. These outliers likely represent individuals with very limited formal education, highlighting a small subset of people with significantly fewer years of schooling compared to the rest of the dataset.

- Capital Gain and Capital Loss: Both the capital gain and capital loss distributions show a high frequency of 0 values, resulting in a large number of outliers. The presence of many 0 values suggests that most individuals in the dataset did not report any capital gains or losses, but the few non-zero values stand out as outliers.

- Hours per Week**:** The boxplot for hours worked per week shows a notable presence of outliers, with a concentration of individuals working less than 30 hours and others working more than 50 hours per week. This suggests a diversity of work schedules, where a significant portion of the dataset either works part-time or works considerably more than the standard 40-hour workweek.

Continuing our analysis, we focus on the categorical features, "workclass", "education", "marital_status", "occupation", "relationship", "race", "sex", "native_country". Let's firstly draw their overall distribution.

Barplots Visualization



Distribution for workclass

Distribution for education

Distribution for marital_status

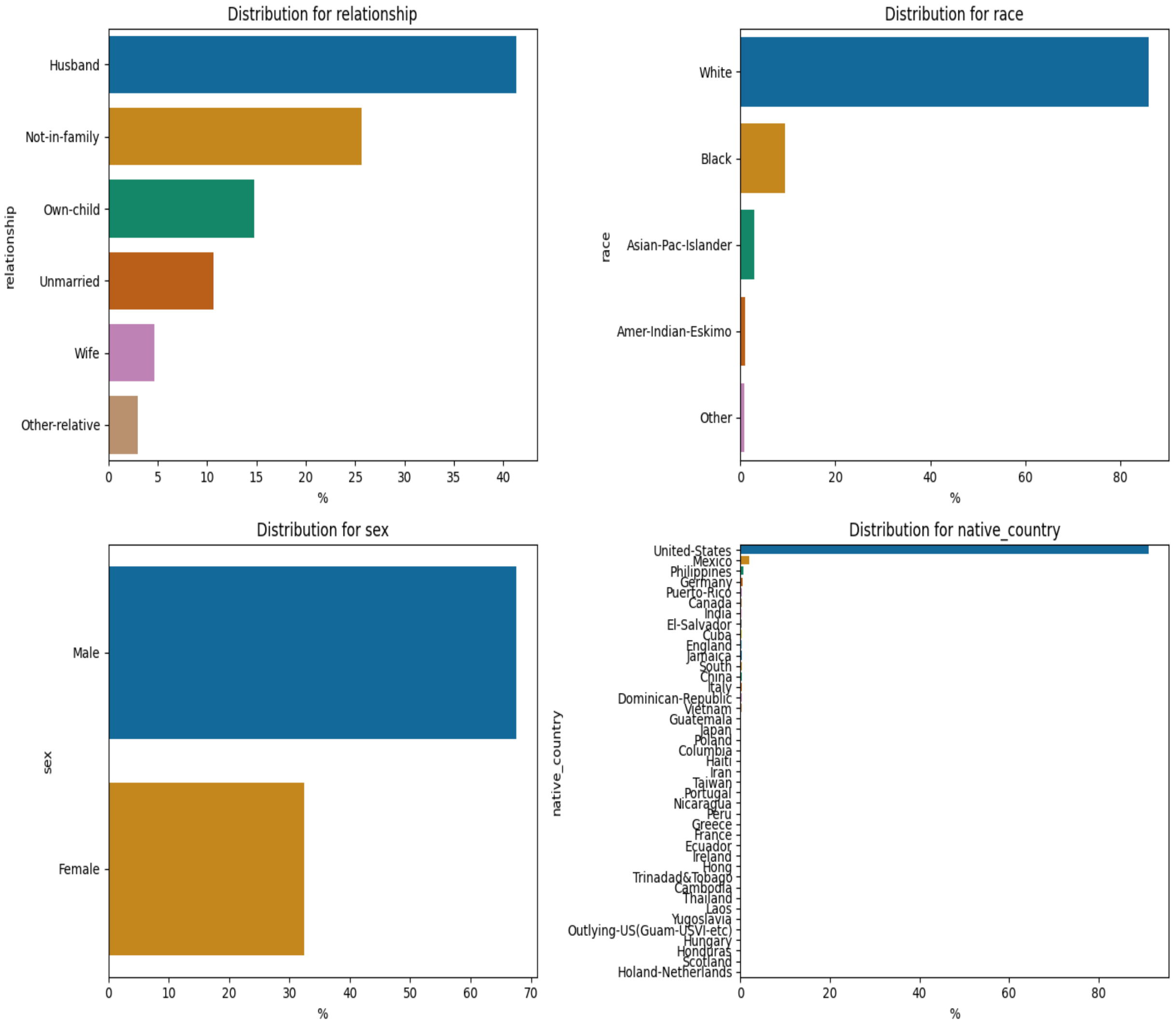
Distribution for occupation

Now, let's take a closer look at each column individually:

Workclass: The workclass distribution shows that over 70% of the individuals in the dataset are employed in the private sector, making it the most common workclass category. This suggests that private-sector jobs dominate the workforce represented in the dataset.

Education: The education column aligns with the data in the "education_num" feature, revealing that the most common educational attainment among the respondents is high school, followed closely by college and bachelor's degrees. This indicates that a significant portion of the dataset has completed secondary education and some higher education.

Occupation: The occupation distribution is quite diverse, with several categories having similar representation. Jobs in craft repair, professional specialties, administrative and clerical roles, executive and managerial positions, sales, and other services each make up approximately 10% of the dataset. This suggests a varied range of occupations within the sample, without any profession overwhelmingly dominating.



Relationship: The relationship status data shows that around 40% of the individuals are classified as husbands, while only about 5% are wives. Additionally, roughly 25% of the individuals are not part of a family unit, indicating that a considerable portion of the dataset is either single or living alone.

Race: The dataset reveals that over 80% of the individuals identify as white. This highlights the racial composition of the dataset, where white individuals make up the vast majority of respondents.

Other columns will be examined in greater detail later in the analysis

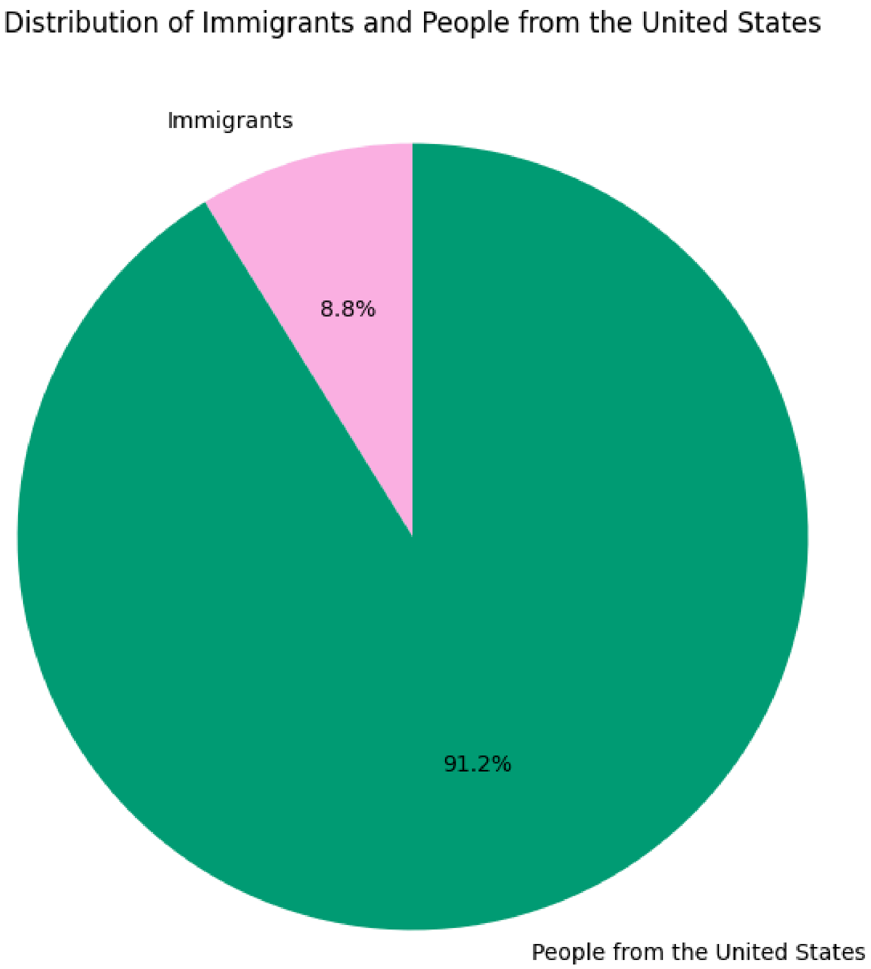
MARITAL_STATUS	COUNT	percentage
married-civ-spouse	14065	46.631523
never-married	9726	32.245872
divorced	4214	13.971222
separated	939	3.135149
widowed	827	2.949559
married-spouse-absent	370	1.292503
married-AF-spouse	21	0.072910

The "marital_status" column offers a representative sampling of the overall population of the United States, with almost the majority of people married, around 32% - never married, and 14% divorced. It will be beneficial to keep this column in our analysis and see the relations with income and marital status.

SEX	COUNT	PERCENTAGE
Male	20380	67.568464
Female	9782	32.431536

The "gender" attribute in the dataset reveals a slight imbalance, with males representing a higher percentage compared to females. While this difference exists, it is not large enough to significantly affect the ability to draw meaningful conclusions or insights from the data. Thus, despite the imbalance, the gender attribute can still provide valuable information for the analysis.

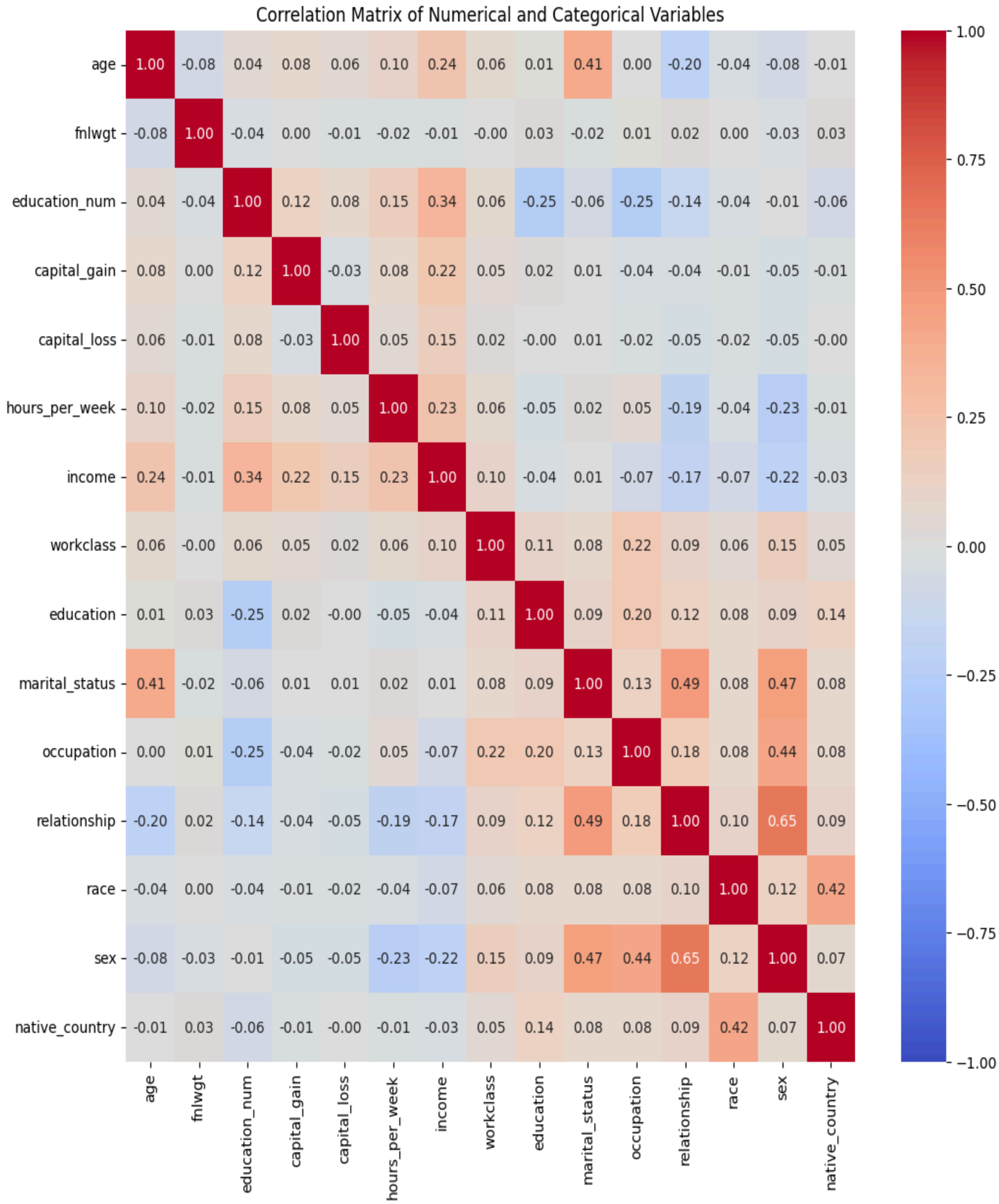
Regarding the "native_country" feature, it reflects a diverse range of countries, offering a representative sample of the U.S. population's geographical makeup. Including this feature in our analysis is beneficial, as it adds depth to understanding the demographic distribution. Additionally, segmenting individuals into two groups—native U.S. citizens and immigrants—may uncover further insights about the differences in characteristics or behaviors between these populations.



The "native_country" feature reveals that U.S. citizens constitute a dominant portion of the dataset, accounting for 91% of all individuals, while immigrants represent a smaller group, making up only 9%. This disproportion highlights the preponderance of U.S. citizens in the sample.

Correlation matrix

In the next step, we will focus on identifying relationships and groupings within our dataset. To begin, we will examine the correlations between the features and the "income" column to identify key relationships and patterns. This will help us understand which factors are most strongly associated with income levels, providing insights for further analysis.



Correlation Matrix of Numerical and Categorical Variables

	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week	income	workclass	education	marital_status	occupation	relationship	race	sex	native_country
age	1.00	-0.08	0.04	0.08	0.06	0.10	0.24	0.06	0.01	0.41	0.00	-0.20	-0.04	-0.08	-0.01
fnlwgt	-0.08	1.00	-0.04	0.00	-0.01	-0.02	-0.01	-0.00	0.03	-0.02	0.01	0.02	0.00	-0.03	0.03
education_num	0.04	-0.04	1.00	0.12	0.08	0.15	0.34	0.06	-0.25	-0.06	-0.25	-0.14	-0.04	-0.01	-0.06
capital_gain	0.08	0.00	0.12	1.00	-0.03	0.08	0.22	0.05	0.02	0.01	-0.04	-0.04	-0.01	-0.05	-0.01
capital_loss	0.06	-0.01	0.08	-0.03	1.00	0.05	0.15	0.02	-0.00	0.01	-0.02	-0.05	-0.02	-0.05	-0.00
hours_per_week	0.10	-0.02	0.15	0.08	0.05	1.00	0.23	0.06	-0.05	0.02	0.05	-0.19	-0.04	-0.23	-0.01
income	0.24	-0.01	0.34	0.22	0.15	0.23	1.00	0.10	-0.04	0.01	-0.07	-0.17	-0.07	-0.22	-0.03
workclass	0.06	-0.00	0.06	0.05	0.02	0.06	0.10	1.00	0.11	0.08	0.22	0.09	0.06	0.15	0.05
education	0.01	0.03	-0.25	0.02	-0.00	-0.05	-0.04	0.11	1.00	0.09	0.20	0.12	0.08	0.09	0.14
marital_status	0.41	-0.02	-0.06	0.01	0.01	0.02	0.01	0.08	0.09	1.00	0.13	0.49	0.08	0.47	0.08
occupation	0.00	0.01	-0.25	-0.04	-0.02	0.05	-0.07	0.22	0.20	0.13	1.00	0.18	0.08	0.44	0.08
relationship	-0.20	0.02	-0.14	-0.04	-0.05	-0.19	-0.17	0.09	0.12	0.49	0.18	1.00	0.10	0.65	0.09
race	-0.04	0.00	-0.04	-0.01	-0.02	-0.04	-0.07	0.06	0.08	0.08	0.08	0.10	1.00	0.12	0.42
sex	-0.08	-0.03	-0.01	-0.05	-0.05	-0.23	-0.22	0.15	0.09	0.47	0.44	0.65	0.12	1.00	0.07
native_country	-0.01	0.03	-0.06	-0.01	-0.00	-0.01	-0.03	0.05	0.14	0.08	0.08	0.09	0.42	0.07	1.00

The heatmap reveals that the correlations between most numerical features, such as "age," "fnlwgt," "education_num," "capital_gain," "capital_loss," and "hours_per_week," are generally low. However, some features show stronger relationships: "age" correlates with "marital_status" and "relationship," "education" is linked to "education_num," and "hours_per_week" is associated with both "sex" and "relationship." "Workclass" and "occupation" also show a correlation, as do "marital_status" with "relationship," "age," and "sex," and "relationship" with "sex." Additionally, "race" correlates with "native_country."

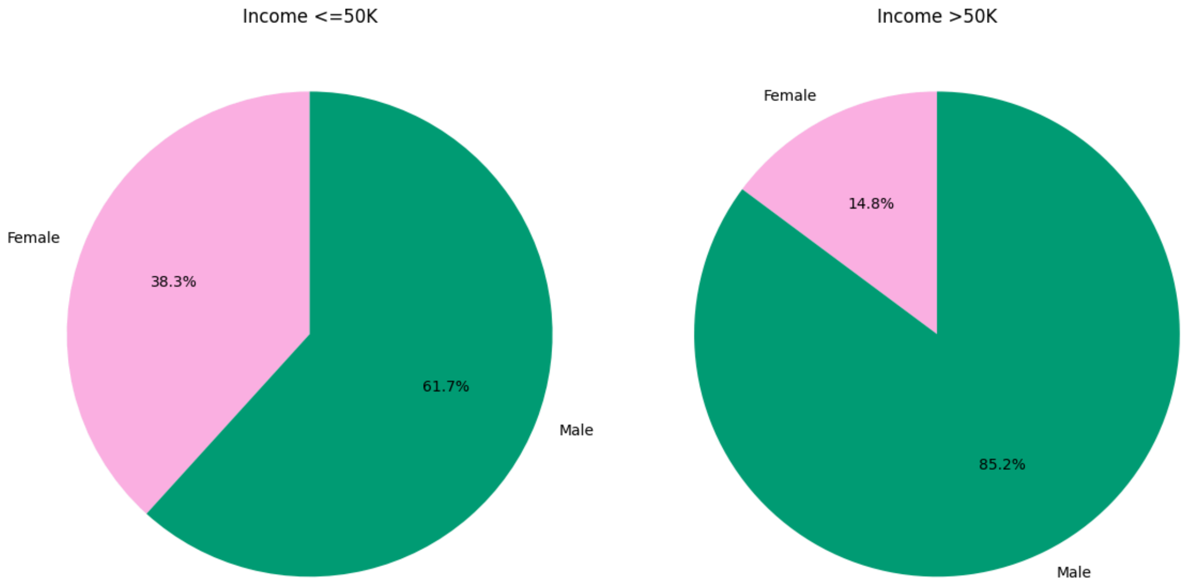
The most positively correlated features with "income" are "education_num," "age," and "hours_per_week." In contrast, "sex" and "relationship" show negative correlations with income.

Note: we have not included the "education" column in the further analysis, opting instead to use the "education_num" column. This choice is due to the fact that "education_num" conveys the same information in a more structured, ordered format. Using "education_num" simplifies the exploratory analysis. However, for the Dashboard, we will use the "education" column, as it is more accessible and easier to interpret for our intended audience.

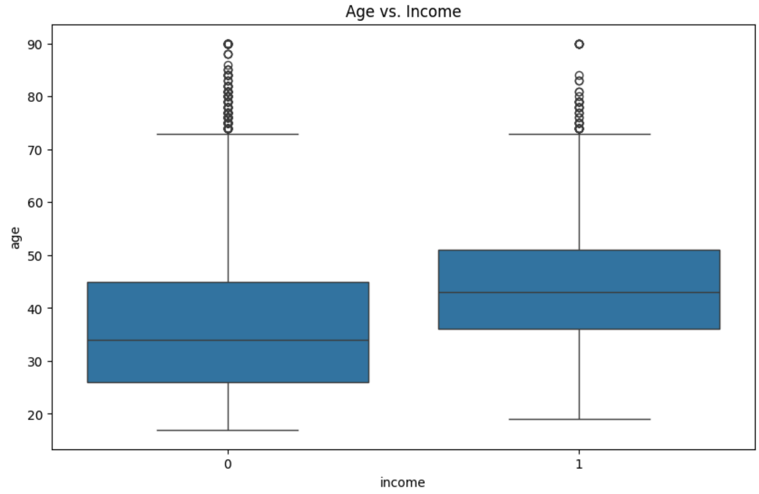
Exploring relationships with the 'Income' variable

We will now focus on the relationships with the "income" column.

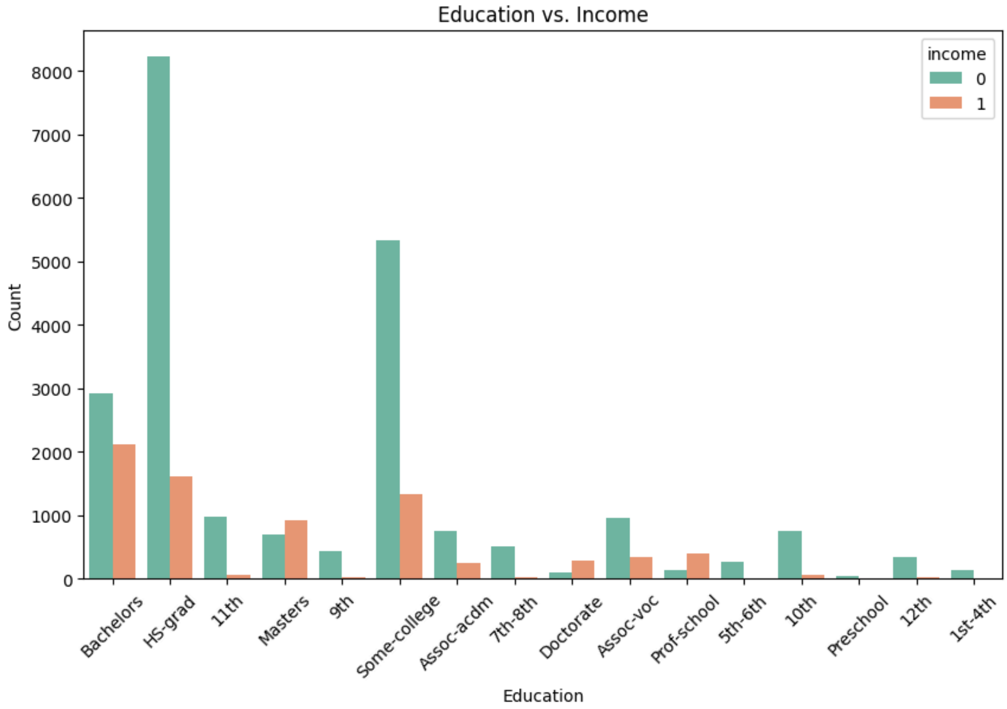
Firstly, we check the income range grouping data by sex.



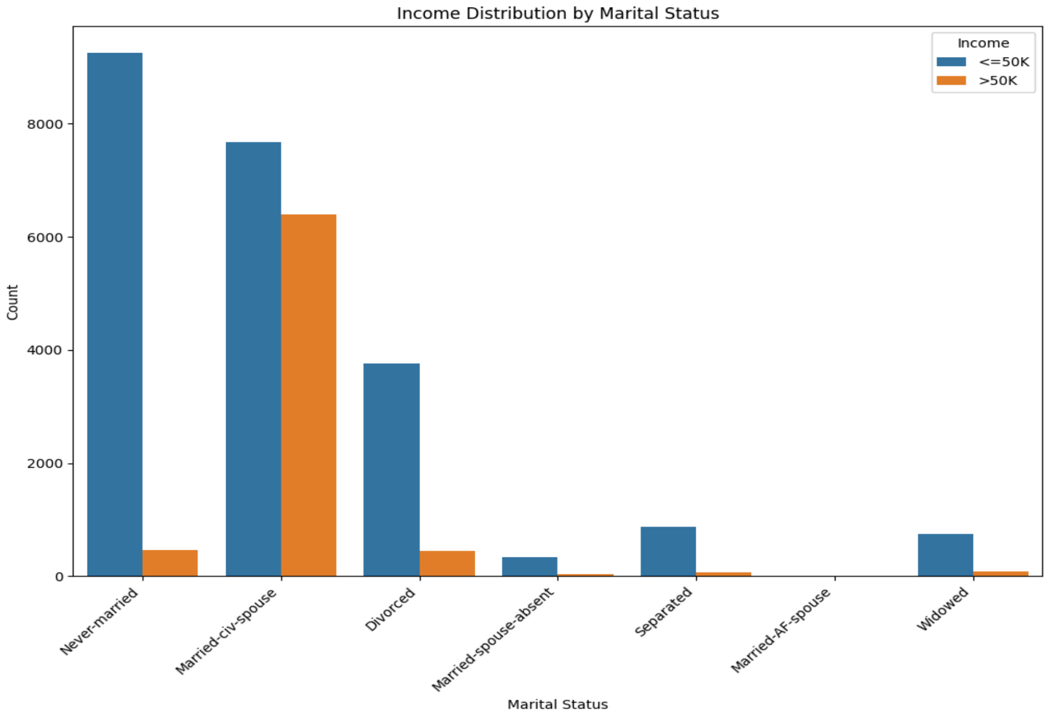
The plot indicates a higher concentration of males in the ">50K" income category, suggesting that gender may play a role in determining income levels. This trend highlights a noticeable gender disparity, with males being more likely to earn above 50K compared to females.



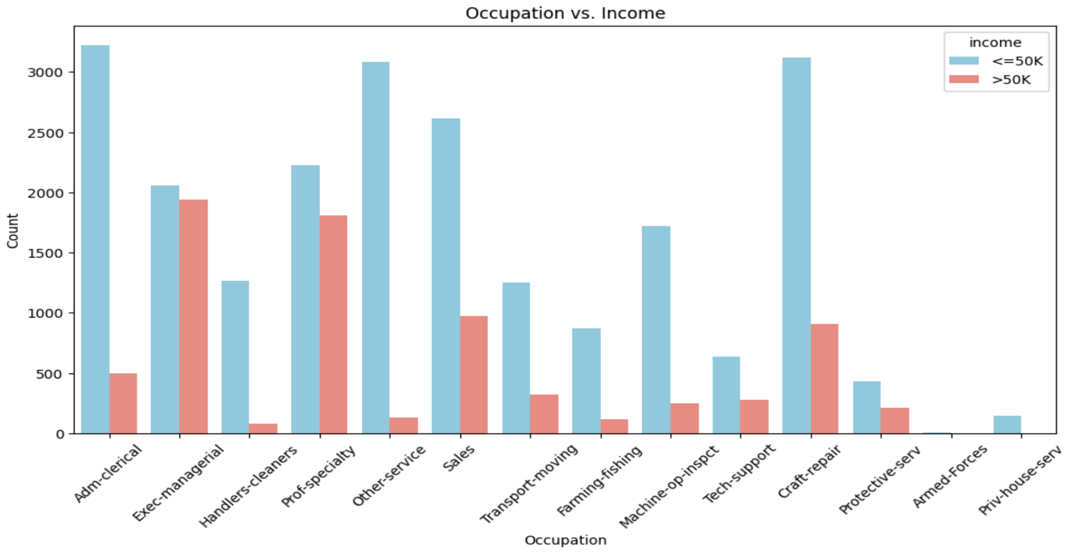
Considering that 1 indicates an income greater than 50K and 0 indicates an income less than 50K, the plot provides a clear visual representation showing that, on average, older individuals tend to earn higher incomes. This is evident from the higher placement of the box in the ">50K" income category for older age groups. Additionally, the plot highlights the presence of outliers, which represent individuals whose incomes are significantly different from the median. These outliers demonstrate the existence of extreme values in the dataset, showing some individuals earn far more or less than the majority. Overall, this visualization offers a concise summary of the relationship between age and income, providing valuable insights into how income varies across different age groups.



The data clearly shows that individuals with higher education levels, such as a bachelor's degree or higher, are more likely to earn ">50K." This trend highlights a positive relationship between education and income. The correlation is further supported by the "education_num" column, which numerically confirms that more years of schooling are associated with higher earning potential.



The plot indicates a potential relationship between marital status and income levels. Specifically, individuals classified as "Married-civ-spouse" appear to have a higher likelihood of earning ">50K, while those labeled as "Never-married" are more frequently found in the lower income category. This suggests that marital status could influence income, with married individuals potentially having more stable financial situations.



We observe that occupations such as "Exec-managerial" and "Prof-specialty" have a higher proportion of individuals earning ">50K," indicating that the type of occupation plays a significant role in determining income levels. This suggests that individuals in higher-ranking or specialized professional roles tend to earn more, further emphasizing the link between occupation and income.

Based on our thorough exploratory data analysis, we have selected the following features to be included in our final dashboard: "age", "sex", "education", "occupation", "native_country", and "marital_status". These features were chosen for their relevance and potential to provide meaningful insights into the dataset's patterns and relationships

Part 2: Audience (top-bottom): Who is your audience?

The project's 2nd task is to describe the audience and to create a persona.

Let's consider an audience that is a government organization or department focused on labor and income policy and analysis, represented by a Government Analyst. These analysts specialize in data analysis, statistics, and policy research centered on labor and income topics. Although they may not be experts in chart design, they primarily work with visualizations on desktop computers or internal government systems. These visualizations are essential tools in their routine evaluations, informing decision-making and supporting policy analysis within their specific areas of responsibility.

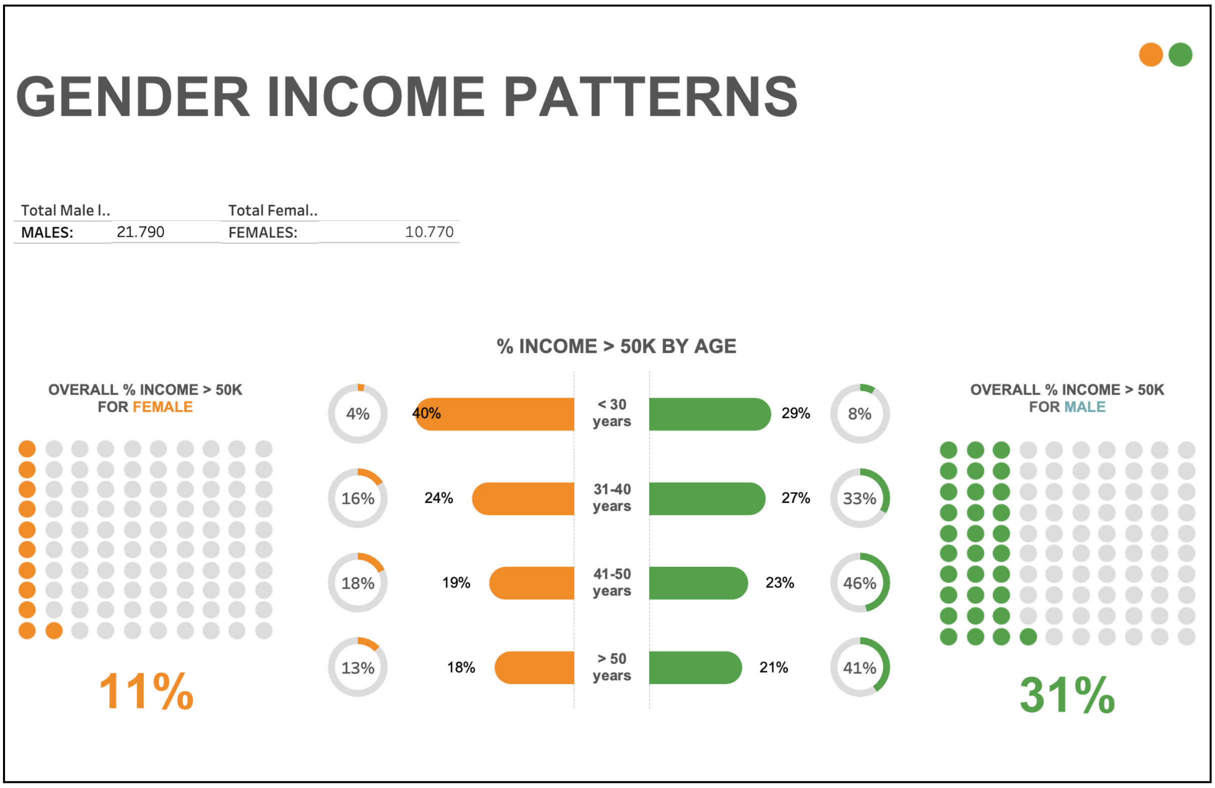
The government analyst's main objective is to understand income disparities, employment patterns, and how factors like education and occupation affect income levels within the population. For our dashboard scenario, we propose the following: The government analyst consults the data visualization while preparing reports on income inequality and analyzing labor market trends.

They want to answer questions like:

- "How does gender or occupation impact income levels?"
- "How does the marital status of each person impact their income?"
- "What is the income distribution among various education levels in the population?"
- "Are there regional variations in income within our country?"

We firstly worked on a draft design to come up with the structure of our dashboard. This "napkin" is a first iteration of our design consisting in a first sketch or a simple approach to the final result.
The visualization type is going to be a dashboard with interactive charts, including bar charts, pie charts, map etc. The dashboard will include two screens (the main dashboard, the detailed dashboard). Both screens will aim to give answers to the previous questions.
We used the following online tableau dashboard as inspiration :https://public.tableau.com/app/profile/simon.evans7317/viz/SportsVizSundayMarchWomensUSOpenFin al/SportsVizSundayMarch2022

This is our first draft for the main dashboard:



This analysis focuses on exploring gender inequality, driven by the significant income differences observed between men and women. As a result, our visualizations will compare these two genders (male and female).

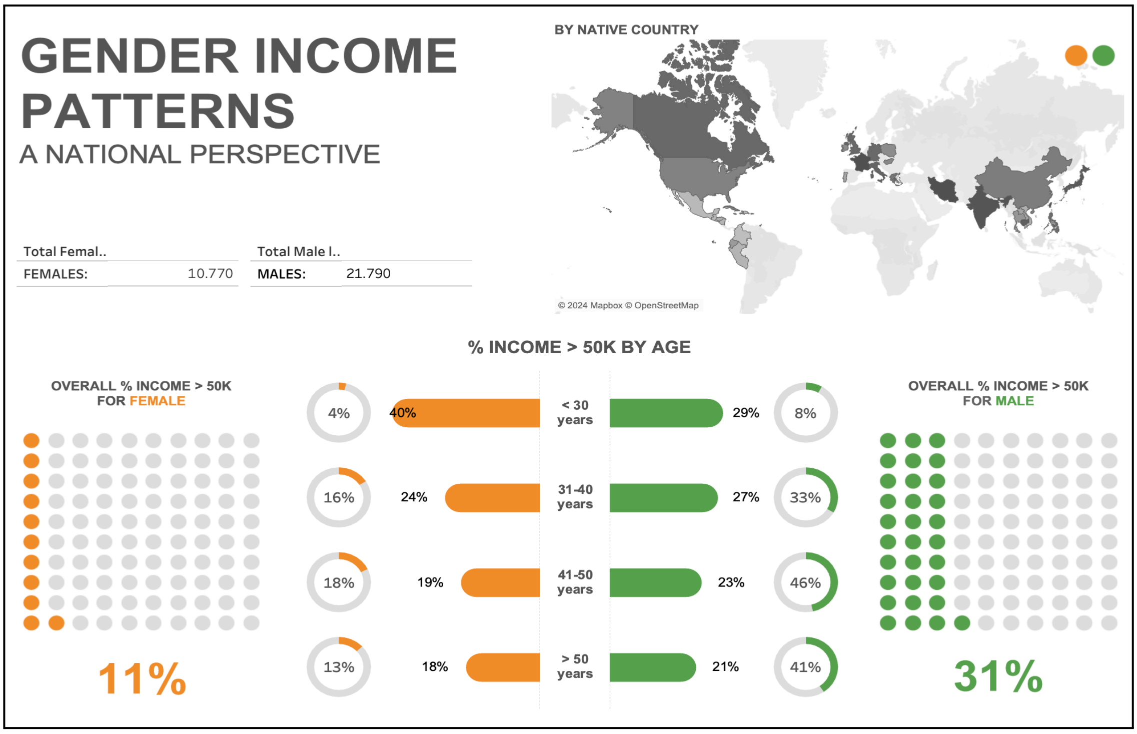
In our approach, we have chosen to represent males in green and females in orange. The dashboard interface includes several key informational elements:

1. Total enumeration of male and female individuals within the dataset's population.
2. The percentage distribution of individuals within each age interval, stratified by gender.
3. The percentage of individuals with an income higher than $50,000, categorized by gender.
4. Representation of the aggregate percentage of individuals with an income exceeding $50,000 for each gender category (Male, Female).

We then realized it would be good to add an additional feature:

5. Implementation of a map visualization to demonstrate the distribution of individuals earning more than $50,000 across various native countries. A filtering mechanism has also been added to be able to do an examination of this data based on gender.

So the draft evolved into this:


GENDER INCOME PATTERNS

A NATIONAL PERSPECTIVE

Total Femal..	
FEMALES:	10.770

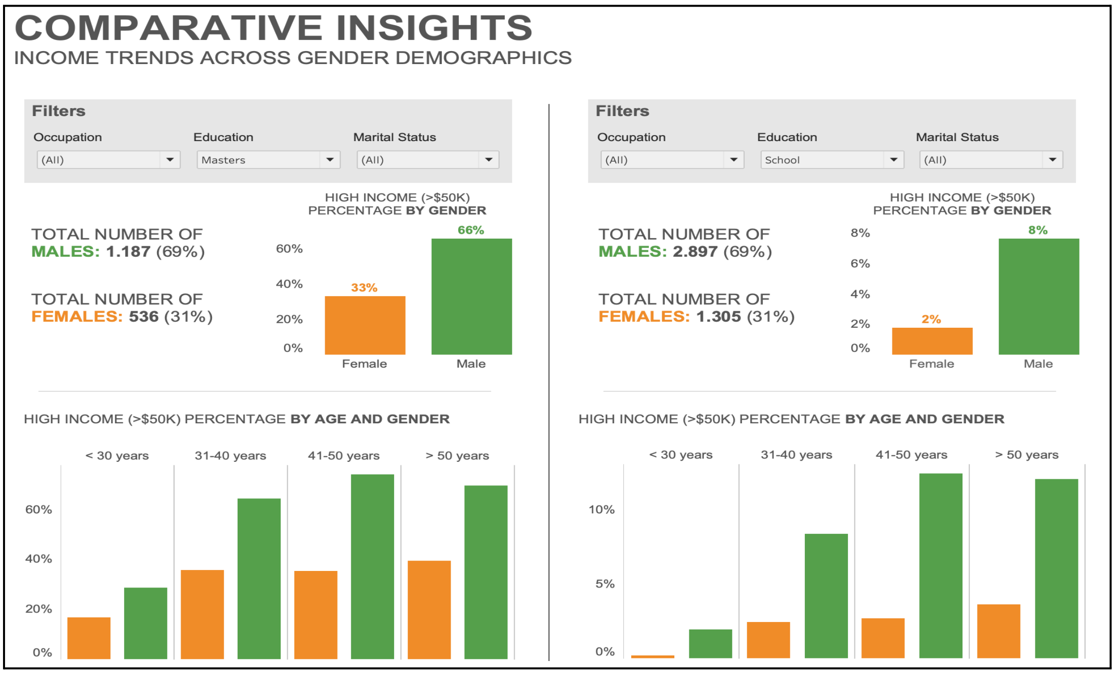
Total Male I..	
MALES:	21.790

For the 2nd dashboard, the objective is to improve interactivity and add more detail.

We have included additional filters, enabling users to access more detailed information. The filters focus on the remaining three features that have not been utilized so far: “occupation”, “education”, and “marital_status”. In this detailed dashboard, we gather information such as:

1. A comparison of occupations between both genders and their impact on income.
2. We use a similar approach to compare various education levels and their influence on income, and the impact of marital status on income.

This is what the 2nd dashboard looks like:


COMPARATIVE INSIGHTS

INCOME TRENDS ACROSS GENDER DEMOGRAPHICS

Filters

Occupation: (All) · Education: Masters · Marital Status: (All)

TOTAL NUMBER OF MALES: 1.187 (69%)

TOTAL NUMBER OF FEMALES: 536 (31%)

Filters

Occupation: (All) · Education: School · Marital Status: (All)

TOTAL NUMBER OF MALES: 2.897 (69%)

TOTAL NUMBER OF FEMALES: 1.305 (31%)

HIGH INCOME (>$50K) PERCENTAGE BY GENDER
60%
40%
20%
0%
33%
66%
Female
Male
HIGH INCOME (>$50K) PERCENTAGE BY AGE AND GENDER
< 30 years
31-40 years
41-50 years
> 50 years
60%
40%
20%
0%
HIGH INCOME (>$50K) PERCENTAGE BY GENDER
8%
6%
4%
2%
0%
2%
8%
Female
Male
HIGH INCOME (>$50K) PERCENTAGE BY AGE AND GENDER
< 30 years
31-40 years
41-50 years
> 50 years
10%
5%
0%

Part 3: Selection of chart and encoding

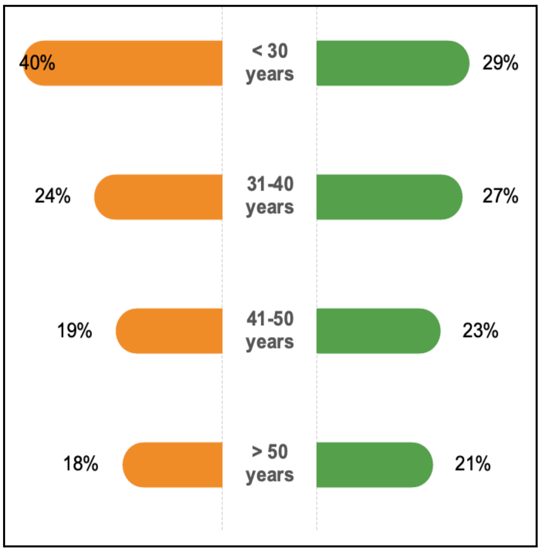
We have to select which charts to use for the dashboard, select the encoding of these ones with justifications.

<u>Main dashboard</u>

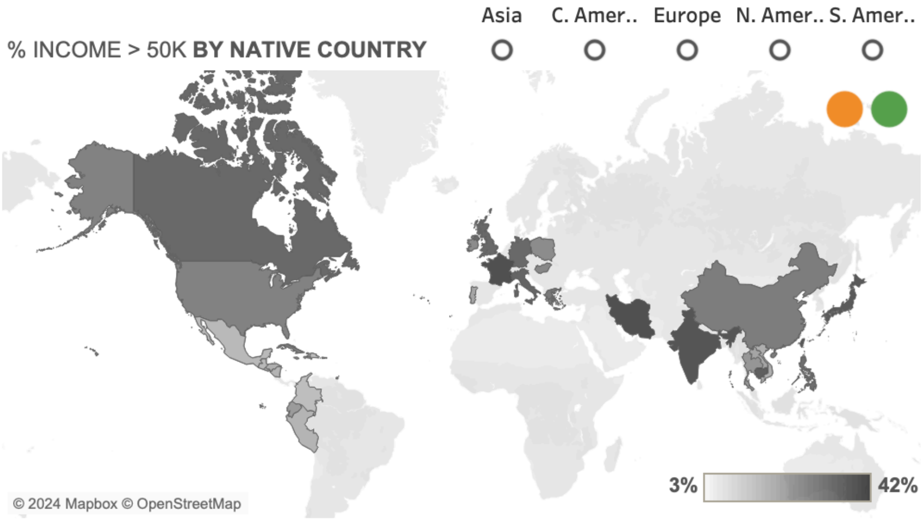
1. Barplot

Encoding: Our presentation uses horizontal bars to illustrate the percentage distribution of individuals across different age groups (<30 years, 31-40 years, 41-50 years, >50 years). Gender distribution is also visually distinguished using two designated colors.

Justification: A barplot effectively displays the relationship between a numerical and a categorical variable. While it may seem like a simple chart, it is likely the most efficient way to present this type of data. For this reason, we have concluded that a bar plot is the best choice to clearly represent and communicate the selected dataset.



2. Chloropleth Map


% INCOME > 50K BY NATIVE COUNTRY
Asia
C. Amer..
Europe
N. Amer..
S. Amer..
© 2024 Mapbox © OpenStreetMap
3%
42%

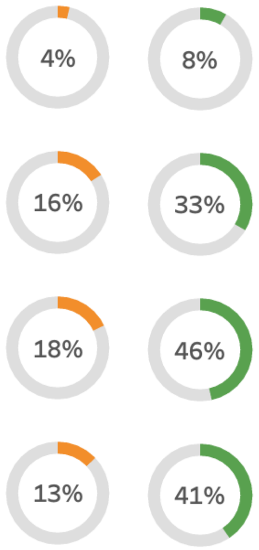
Encoding: The map displays the countries of origin for individuals working in the U.S. with incomes over $50,000. This dataset is from 1994, and countries that no longer exist, like Yugoslavia, have been omitted for this project. To improve clarity and analytical depth, we've added two main filters: (1) a gender filter to separate data by male and female categories, and (2) a geographical filter to sort data by continent or region (Asia, Central America, Europe, North America, South America). Also, users can hover over each country to view detailed information, including the country's name, the percentage of individuals from that country earning over $50,000, and the total population from that country.

Justification: Geographical data is more intuitive when displayed on a map, as it allows users to quickly grasp regional income trends. With the addition of filters for gender and continent/region, the map provides multidimensional insights, offering a deeper understanding of income distribution. Overall, the map acts as a valuable tool for clear communication, data-driven decision-making, and identifying patterns in income distribution.

3. Doughnut chart

Encoding: Using a doughnut chart, we visually depict the percentage distribution of individuals earning over $50,000 for each gender. Females are represented in orange, while males are shown in green. These charts are connected to the earlier bar plot, creating a comparative framework that contrasts each percentage with its corresponding age group.

Justification: A doughnut plot is used to visually represent the values of multiple entities, especially to show proportions that total 100%. This choice was made to clearly illustrate the distribution between the two genders within each age group and to highlight the number of individuals earning over $50,000.

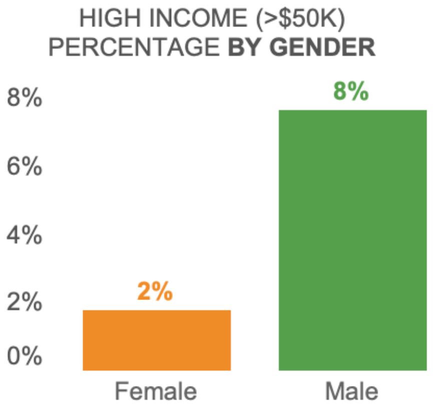


<u>Second dashboard</u>

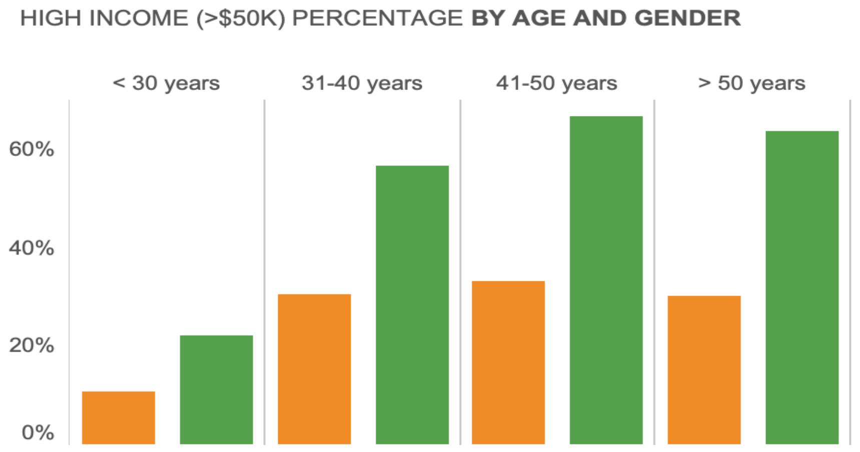
1. Barplot

Encoding: This barplot effectively illustrates the disparity between the two genders across any chosen feature. Consequently, it allows us to draw conclusions regarding the percentage of individuals with an income surpassing $50,000. For instance, selecting any education/occupation/marital status option dynamically alters the bar plot, facilitating a comprehensive analysis.

Justification: We determined that employing the barplot in this section is more comprehensive than utilizing pie charts, as the percentages across the two genders do not necessarily add up to 100%.



2. Barplot



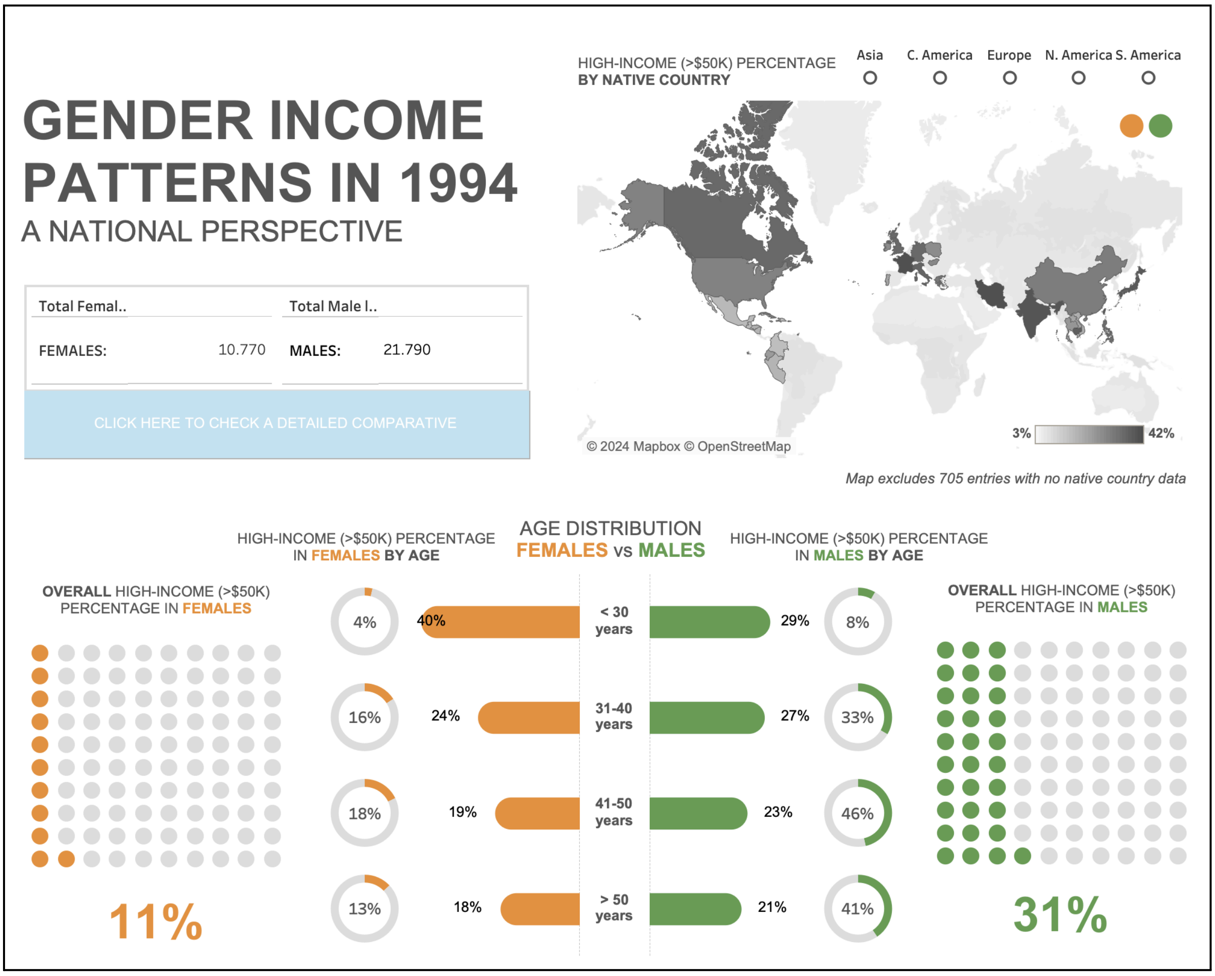
Encoding: In this section, our objective is to distill the same information as presented in the main dashboard, where we segmented the dataset into age intervals (<30 years, 31-40 years, 41-50 years, >50 years) and conducted a comparative analysis between males and females. However, in this segment, we focus on examining the percentages of females and males with an income exceeding $50,000 within each age group. This is achieved through filtering based on education, occupation, and marital status. Consequently, we provide users with the flexibility to utilize the filtering options and draw conclusions in any manner they find meaningful.

Justification: To ensure a clear and user-friendly interface, we selected a bar plot for the dashboard. This decision supports our objective of offering a straightforward and visually accessible tool that allows users to easily navigate and derive valuable insights.

Part 4: Implementation

For the final delivery, we made several improvements to the dashboard by incorporating various adjustments.

1. Improvements have been made to the dashboard's clarity and ease of understanding by refining the titles of each chart, as well as the main title, to provide a more accurate presentation of the information.
2. We sorted out the size issues in the bar charts to ensure accurate and consistent representation of the data.
3. We introduced a button that allows navigation to the second dashboard.



GENDER INCOME PATTERNS IN 1994

A NATIONAL PERSPECTIVE

Total Femal..		Total Male I..	
FEMALES:	10.770	MALES:	21.790

CLICK HERE TO CHECK A DETAILED COMPARATIVE

HIGH-INCOME (>$50K) PERCENTAGE
BY NATIVE COUNTRY
Asia
C. America
Europe
N. America
S. America
3%
42%
© 2024 Mapbox © OpenStreetMap
Map excludes 705 entries with no native country data
HIGH-INCOME (>$50K) PERCENTAGE IN FEMALES BY AGE
AGE DISTRIBUTION FEMALES vs MALES
HIGH-INCOME (>$50K) PERCENTAGE IN MALES BY AGE
OVERALL HIGH-INCOME (>$50K) PERCENTAGE IN FEMALES
OVERALL HIGH-INCOME (>$50K) PERCENTAGE IN MALES
4%
40%
< 30 years
29%
8%
16%
24%
31-40 years
27%
33%
18%
19%
41-50 years
23%
46%
13%
18%
> 50 years
21%
41%
11%
31%

In conclusion, our dashboards enable us to effectively respond to any remaining questions from our users, providing thorough solutions to their previous inquiries.

- It is important to highlight that gender plays a significant role in income, with women earning much less than men. Additionally, specific occupations, like administrative-clerical and craft-repair roles, show considerable income gaps, highlighting the widespread impact of gender disparities across different job sectors.
- Income distribution is clearly shaped by educational attainment, with individuals who have higher levels of education generally earning more. This relationship highlights the significant role education plays in determining income levels.
- Marital status seems to have a significant impact on income, as married individuals tend to earn significantly more than those who are divorced or unmarried.
- We have observed that an individual's native country does not have a noticeable effect on their income.

Overall, our designed dashboard features a well-organized layout that emphasizes clarity and simplicity on both screens. By focusing on these elements, we aim to provide a smooth user experience and make it easy for users to draw the desired conclusions.