Presentation and Visualization

Delivery 1: Dashboard

20th November 2023 – Dafni Tziakouri, Victor Fayos, Alejandro Vara, Pere Blanch

First of all, we selected the **"Adults"** dataset from Census ("Census Income"), as the dataset we will use during this course to create our dashboards. Census is an official website of the United States government, and the data is about determine whether a person makes over 50K a year or less. Furthermore, it is important to acknowledge that the data utilized in this study was extracted by Barry Becker in 1994. Consequently, certain aspects may contain dated information; for instance, our dataset includes references to Yugoslavia, a geopolitical entity that no longer exists as a country.

Dataset link: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

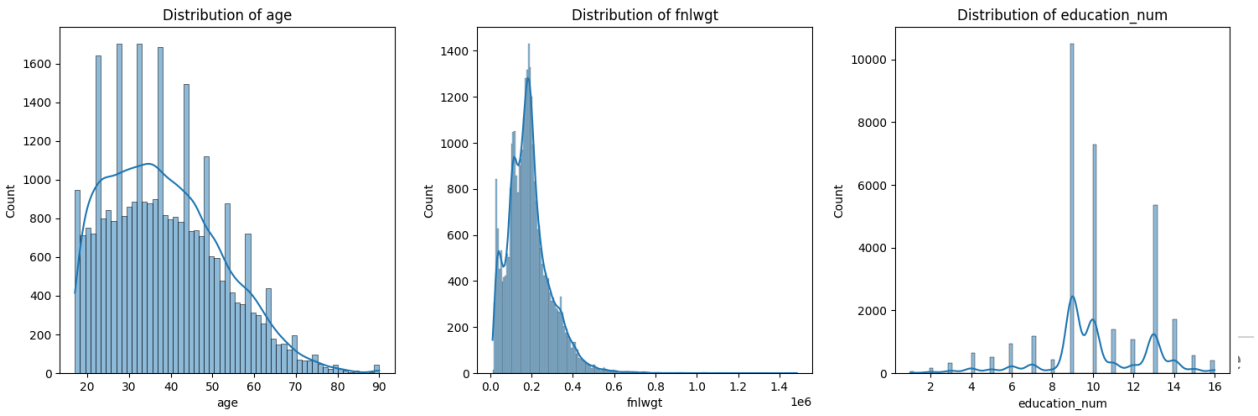
Dashboard link: https://public.tableau.com/app/profile/pere.blanch.ramirez/viz/dashboard_pv/Dashboardmain?publish=yes (*Note*: In the public tableau some lines are visible, which we do not have them in our local dashboards)

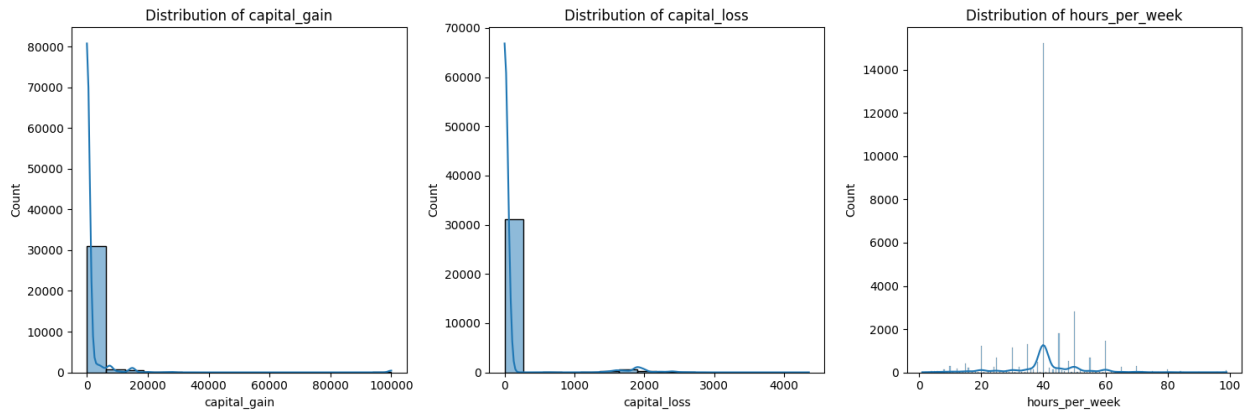
Part 1: Exploratory Data Analysis (EDA): Data (bottom-up): Study and describe the data + explore relationships.

Our first task in this project is to describe the data included in the chosen dataset and explore some relationships between the data.

The type of our data is divided in two types: categorical data and numerical data. The categorical data includes features like "workclass," "education," "marital_status," "occupation," "relationship," "race," "sex," and "native_country" and the numerical data includes features like "age," "fnlwgt" (final weight), "education_num," "capital_gain," "capital_loss," and "hours_per_week". Therefore, we have 14 features and we would like to choose 5-6, to create our dashboards.

The dataset does not include explicit temporal information, because it is not a time-series dataset but it does have a geographical nature. The "native_country" feature represent the country of origin of individuals and it can be used to explore geographic patterns in the data. In general, the range, units, and precision of numerical features vary depending on the specific feature. Furthermore, the dataset can be considered as static and does not change over time. It represents a snapshot of the data at a specific point in time.

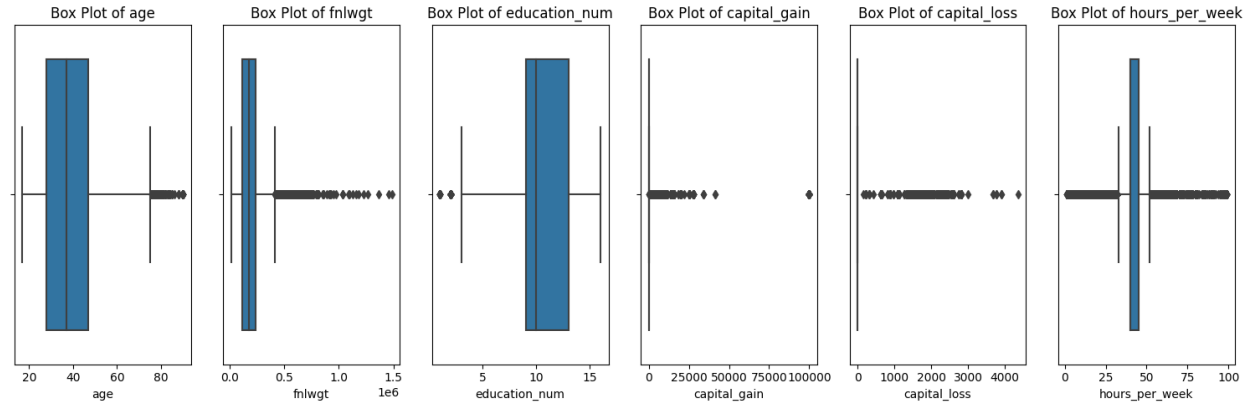




Therefore, we obtained the following conclusions:

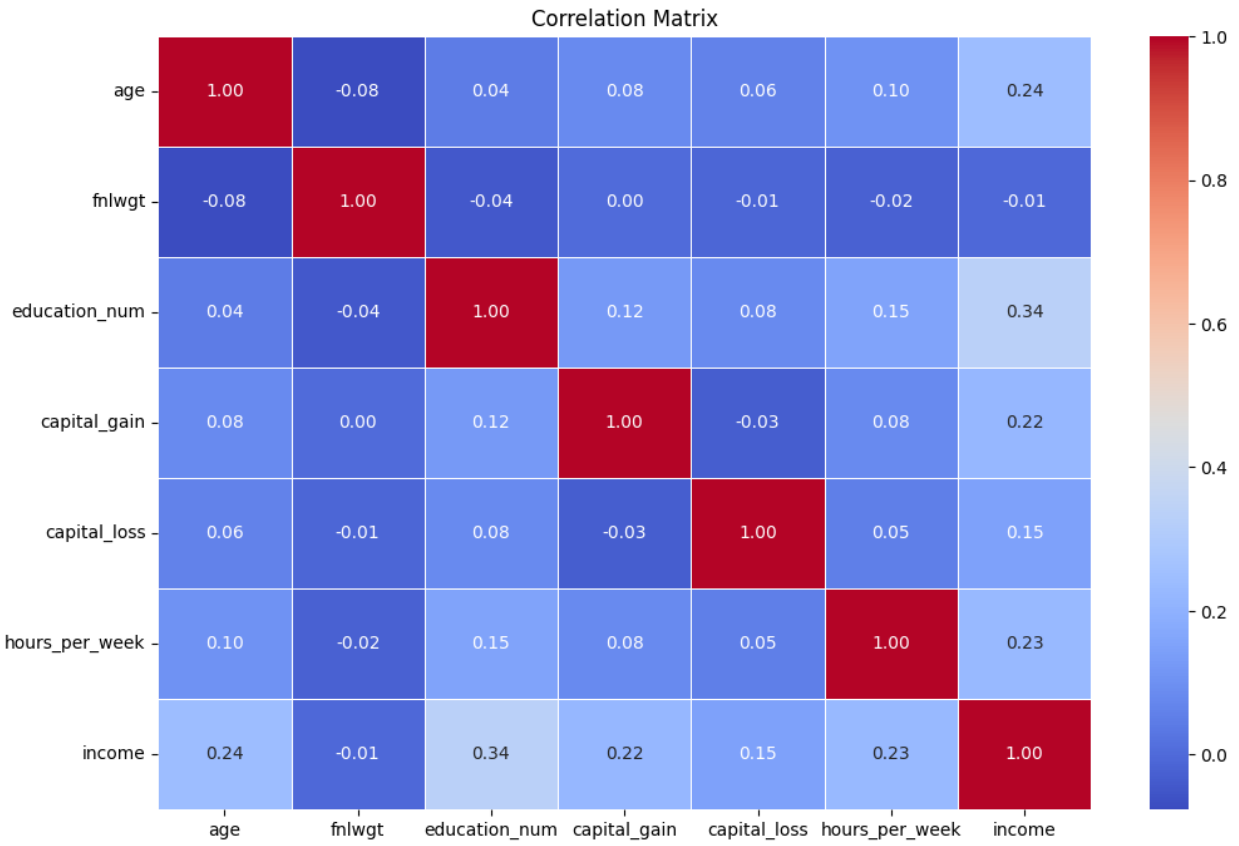
- When examining the “age” distribution, it becomes evident that it is slightly right-skewed, with a peak occurring in the mid-30s, indicating that a substantial proportion of the dataset comprises individuals in their 30s. This skewness is further supported by the density plot, which offers a smoother representation of the data distribution.
- The distribution of “hours_per_week” seems to follow a roughly normal distribution, with a prominent peak at around 40 hours per week, a typical representation of full-time employment. This normality is affirmed by the density plot, which exhibits a distinct peak at 40 hours per week.

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 “age” boxplot shows a few outliers on the high end, indicating individuals who are significantly older than the majority of the dataset. These outliers suggest a presence of older individuals.
- The “hours_per_week” boxplot doesn't reveal clear outliers, suggesting that the distribution of hours worked per week is relatively uniform without extreme values.

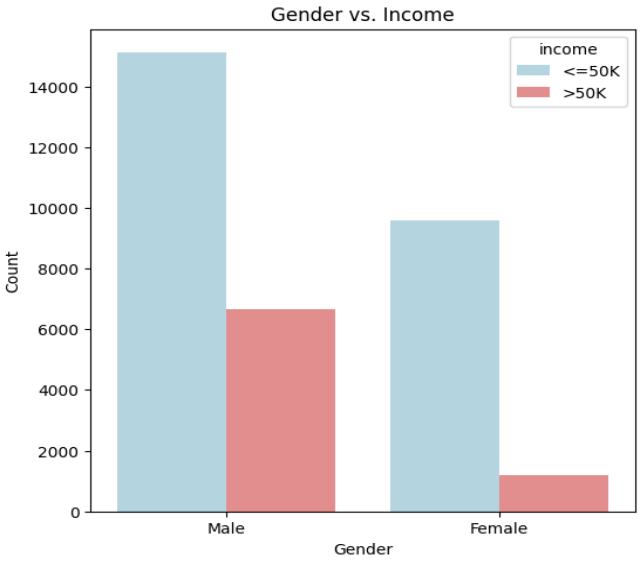
Let us proceed by generating various plots to examine the correlations among our features:



As evident from the preceding heatmap, correlations among the attributes "age", "fnlwgt", "education_num", "capital_gain", "capital_loss" and "hours_per_week" are nearly negligible.

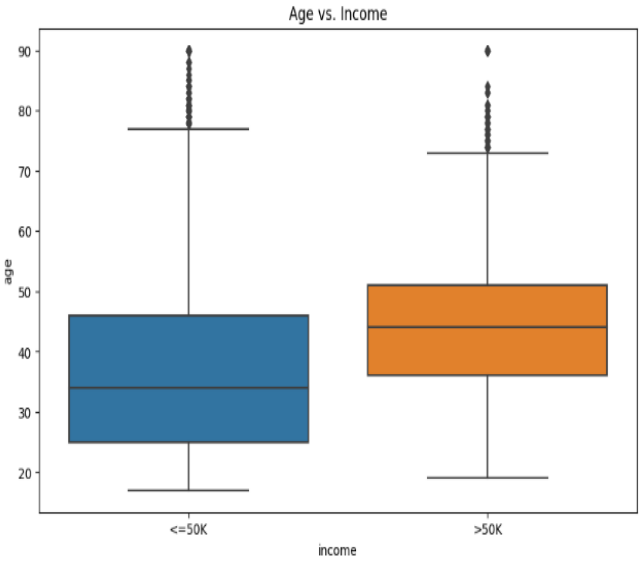
However, when assessing the correlation with the "income" column, the three most positively correlated attributes, in descending order, are "education_num", "age" and "hours_per_week". Additionally, given that the correlation with "fnlwgt" could be considered null, this provides a compelling rationale for excluding it from further consideration.

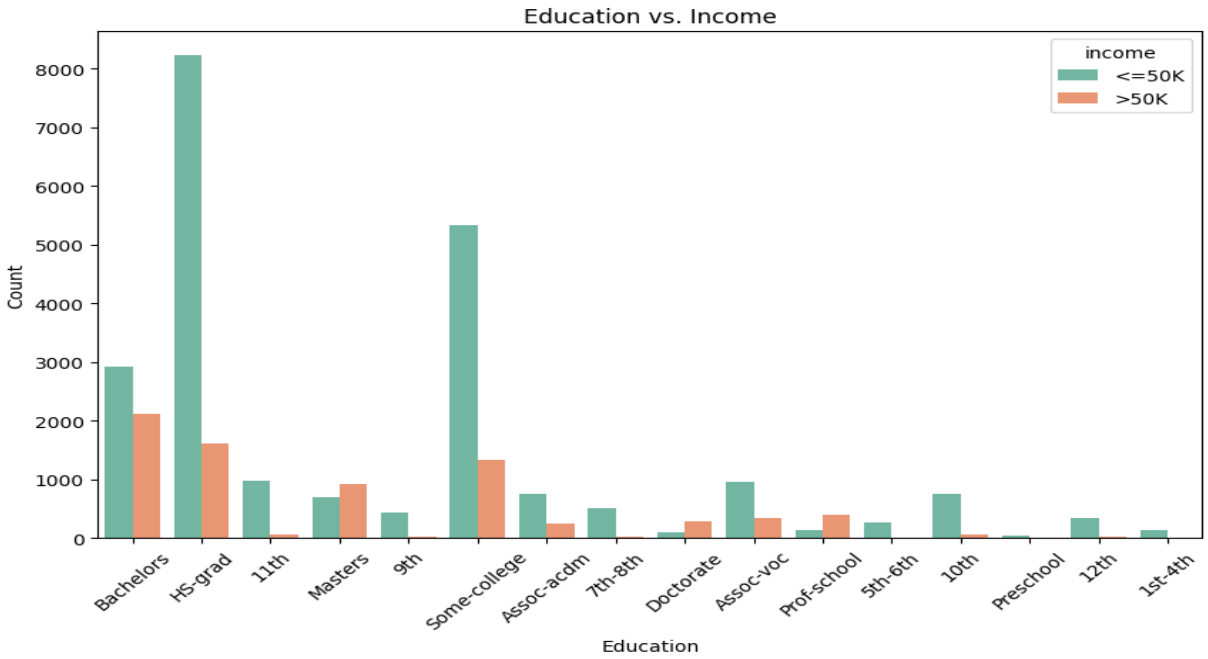
Note: Up to this point, we have not considered the "education" column in our analysis. This decision is based on the presence of the "education_num" column, which encapsulates the same information but with an ordered relationship. Consequently, utilizing "education_num" facilitates a more straightforward exploratory analysis. However, "education" will be the one considered in the Dashboard as it is more easily readable for our target audience.



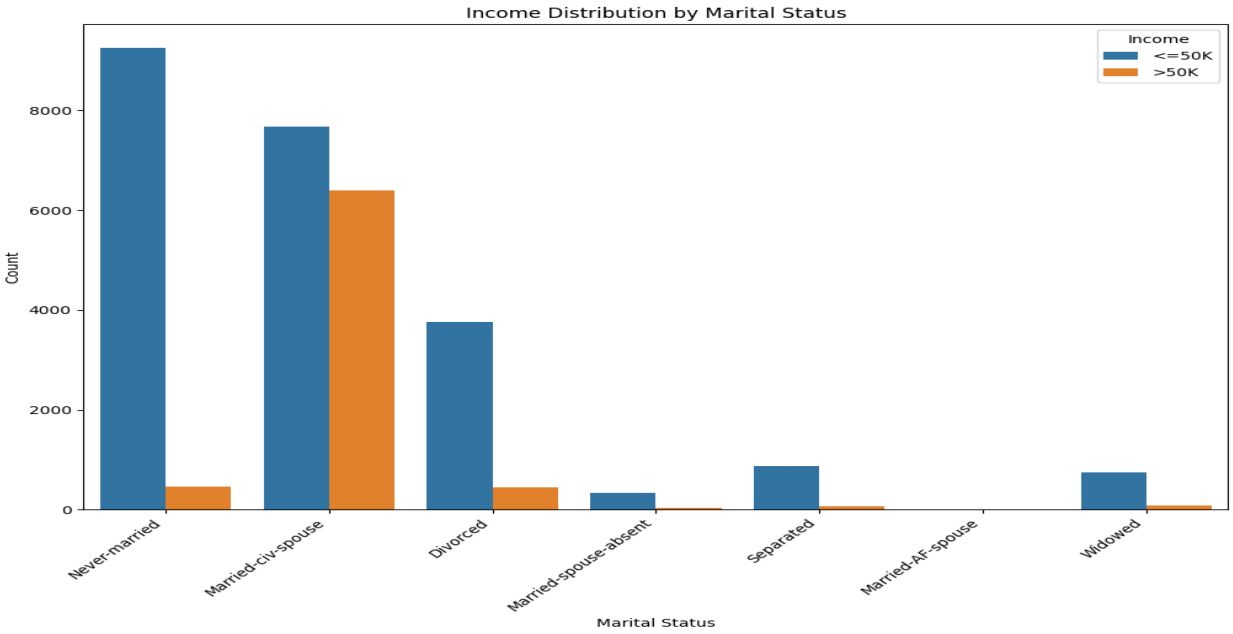
Evidently, there is a noticeable prevalence of the ">50K" income category among males, suggesting a discernible gender-based income disparity. This observation prompts further investigation into the factors contributing to this divergence, thereby enhancing our understanding of the underlying dynamics.

It clearly illustrates that, on average, older individuals tend to have higher incomes, as demonstrated by the higher box position for the ">50K" income category. The plot also highlights the presence of outliers, indicating individuals with incomes significantly different from the median. This visualization succinctly summarizes the income dynamics with respect to age and provides valuable insights into the dataset's income distribution.

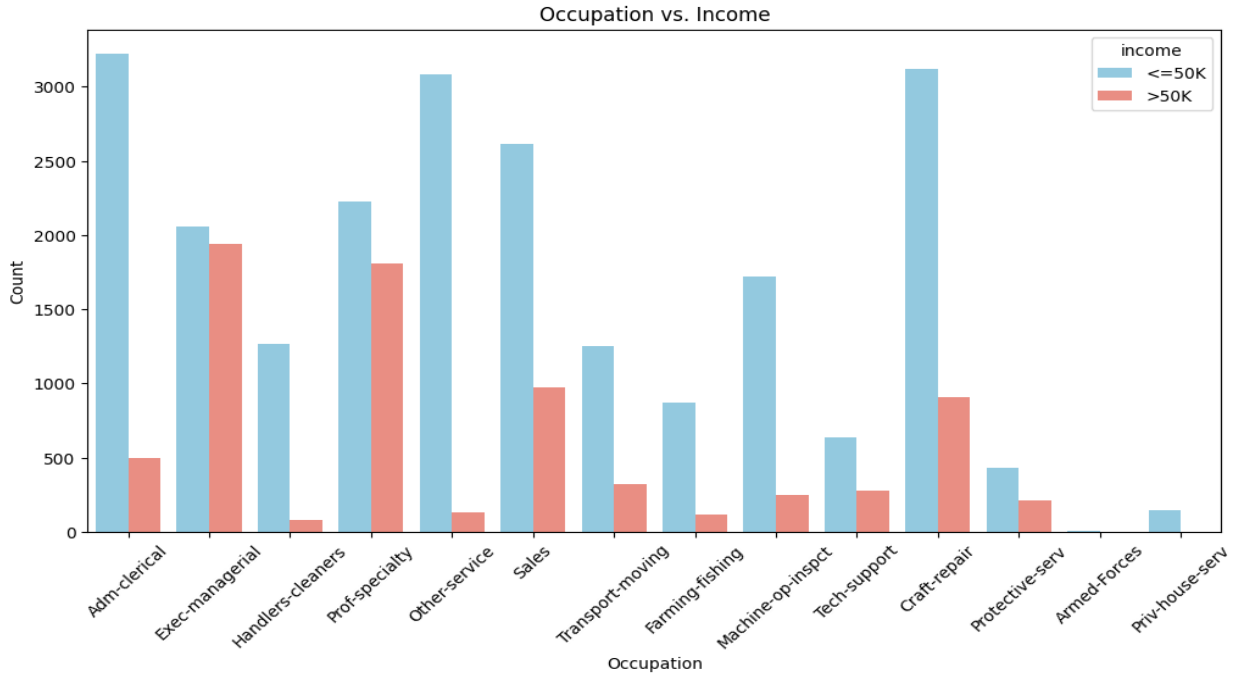




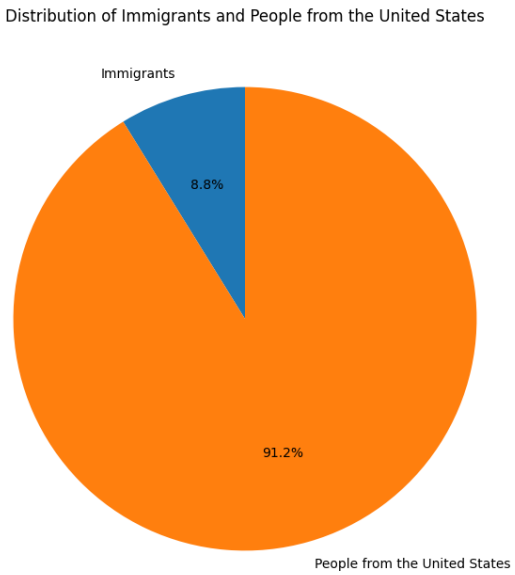
Clearly, there is a discernible pattern indicating that individuals with advanced education levels, specifically those with a bachelor's degree or higher, exhibit a greater probability of earning ">50K." This observation underscores a positive correlation between educational attainment and income, emphasizing the significance of higher education in influencing income levels.



The plot, suggests that marital status may be associated with income levels, with "Married-civ-spouse" having a higher likelihood of higher incomes, while "Never-married" individuals are more prevalent in the lower income category.



Notably, certain occupations, such as "Exec-managerial" and "Prof-specialty," tend to have a higher proportion of individuals earning ">50K," indicating the potential influence of occupation on income.



The data presented in the "native_country" feature appears to be a representative sampling of the overall population of the United States. Consequently, it seems prudent to consider this column in our analysis.

After all this exploratory data analysis, we decided to select the following features as the final ones to create our dashboard: "age", "sex", "education", "occupation", "native_country" , "marital_status".

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

The second task of this project is to describe the audience and to create a persona.

So, let's assume the audience is a government organization or department responsible for labor and income-related policies and analysis (Government Analyst). Their expertise lies in data analysis, statistics, and policy research, with a specific focus on matters related to labor and income, but not necessarily be an expert in chart design. In their work, the Government Analyst primarily engages with visualizations on desktop computers or internal government systems. These visualizations serve as a crucial resource for their regular consultations, supporting decision-making processes and policy analysis within their domain of responsibility.

The government analyst's primary goal is to gain insights into income disparities, employment trends, and the impact of education, occupation, and other factors on income levels within the population. We propose the following scenario to use for our dashboard: The government analyst accesses the data visualization when preparing reports on income inequality and labor market analysis.

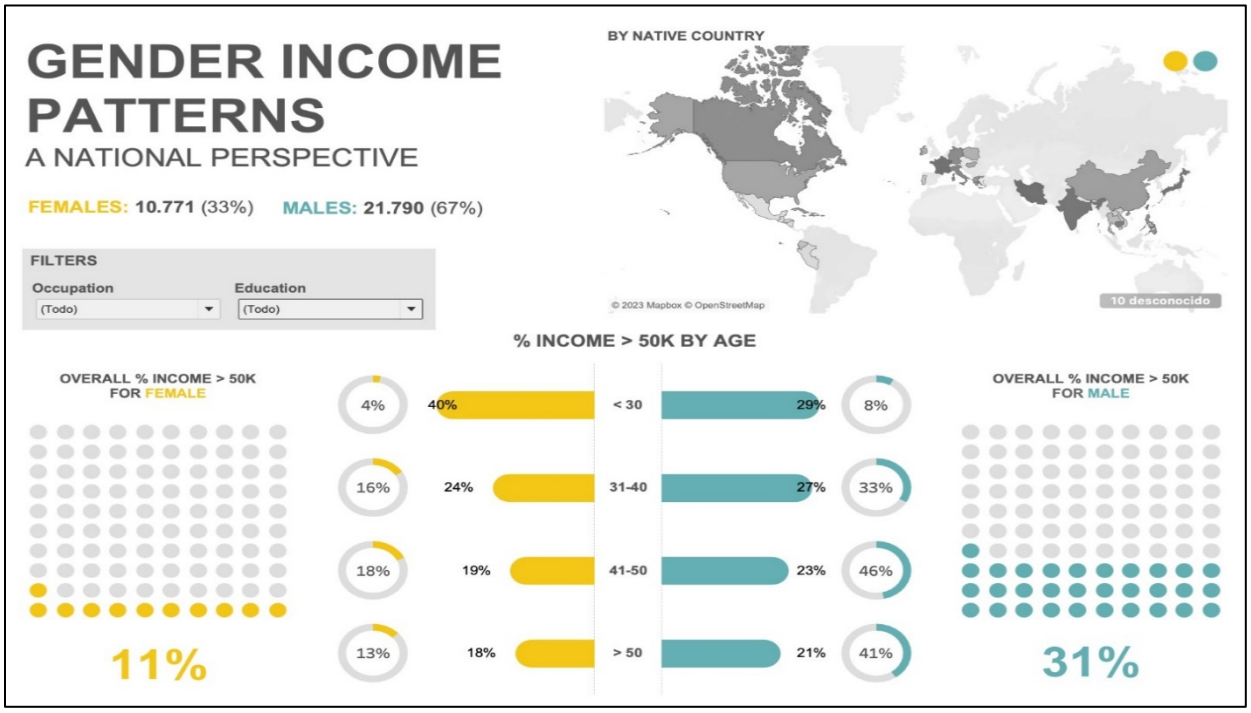
They want to answer questions such as:

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

We proceed with the development of a “napkin” design to articulate the structure of our dashboard effectively. This “napkin” will be an initial 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 multiple interactive charts, including bar charts, pie charts, map etc. We are expecting our dashboard to include two screens (the main dashboard, the detailed dashboard). In addition, both screens must give answers to the previous questions.
For that, we got inspired from the following online tableau dashboard: https://public.tableau.com/app/profile/simon.evans7317/viz/SportsVizSundayMarchWomensUSOpenFinal/SportsVizSundayMarch2022

Here is our first trial for the main dashboard:

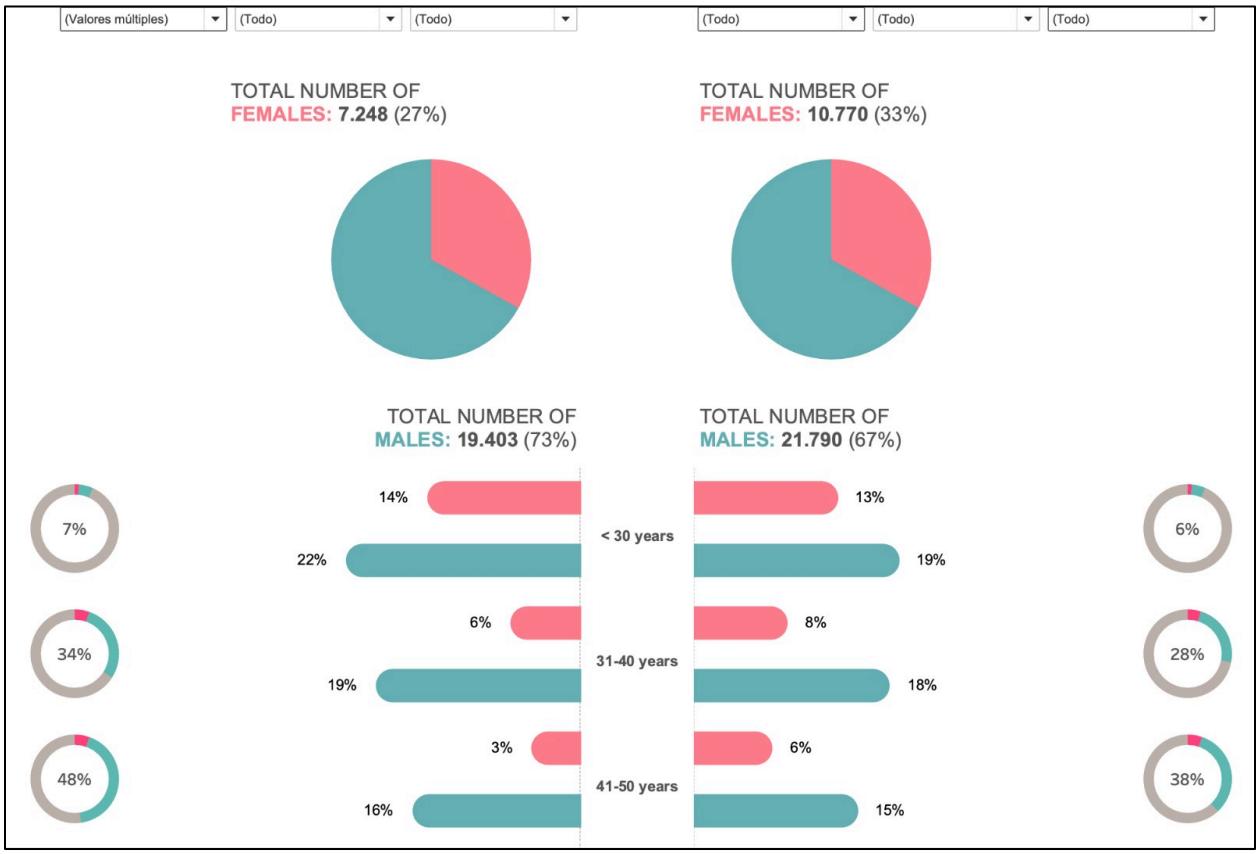


Our emphasis in this analysis is directed towards examining gender inequality, prompted by the notable and substantial income disparities observed between genders. Therefore our visualizations will be in comparison between the two genders (male, female).

In our visualization methodology, we have opted to represent males with the color blue and females with the color yellow (the colors are still under discussion, so we will select the best pallet for our dashboard). Within the dashboard interface, several informative elements are presented:

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).
5. Implementation of a map visualization to demonstrate the distribution of individuals earning more than $50,000 across various native countries. Additionally, a filtering mechanism has been incorporated to facilitate an examination of this data based on gender.
6. Addition of filters so the user can select specific education levels, occupation categories, or geographic regions to view detailed insights and comparisons.

In designing the second dashboard, our objective is to augment interactivity and enhance detail. Here is our first trial for the second dashboard:



Our intention is to maintain a consistent structure with the main dashboard to achieve cohesion between the two screens. Furthermore, in the second dashboard, we have incorporated additional filters, allowing users to access more detailed information. This enables users to draw conclusions on more specific topics. The filters will focus on the remaining three features that have not been utilized this far: "education," "occupation," and "marital_status". It's worth mentioning that we are experimenting with various color combinations in our pursuit of the optimal outcome. In this detailed dashboard, we gather information such as:

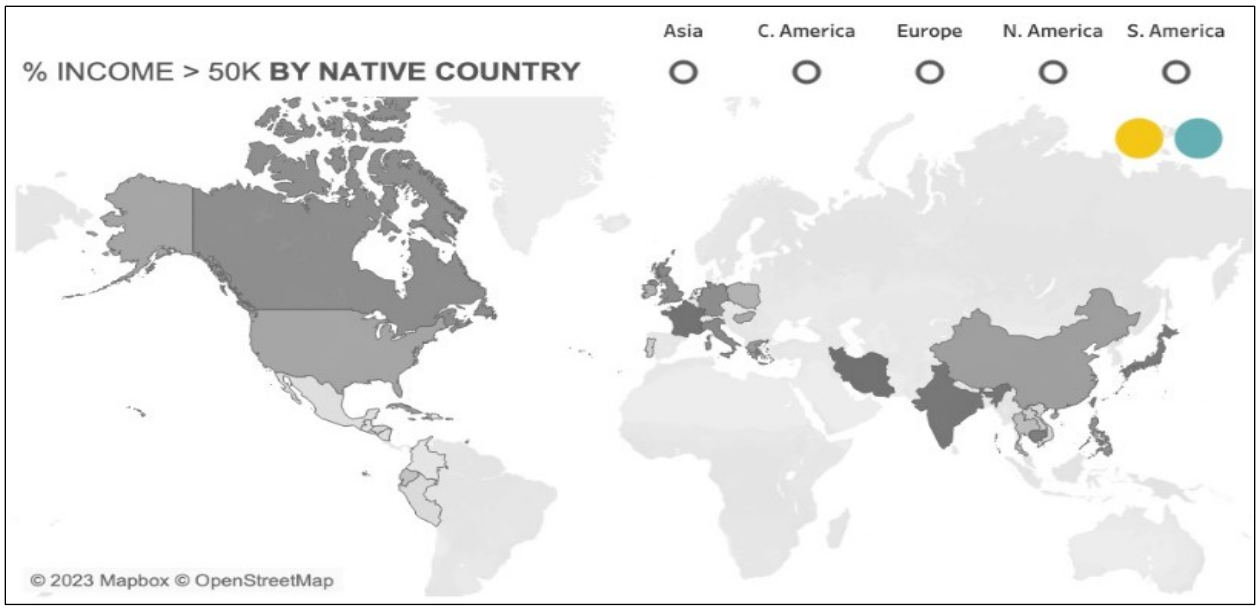
1. A comparison of occupations between both genders and their impact on income. Initially, we contemplated presenting this information through pie charts, but we are still evaluating this approach.
2. We will employ a similar approach to compare various education levels and their influence on income, as well as the influence of the marital status on income.

We are currently evaluating the suitability of these particular plots to optimize the overall structure for our desired outcome.

Part 3: Selection of chart and encoding

In order to achieve the objectives we stated previously, we need to select which charts we want to use for our dashboard, select the encoding of these ones and justify these selections. Let's discuss the decisions taken for each chart and each screen of our dashboard.

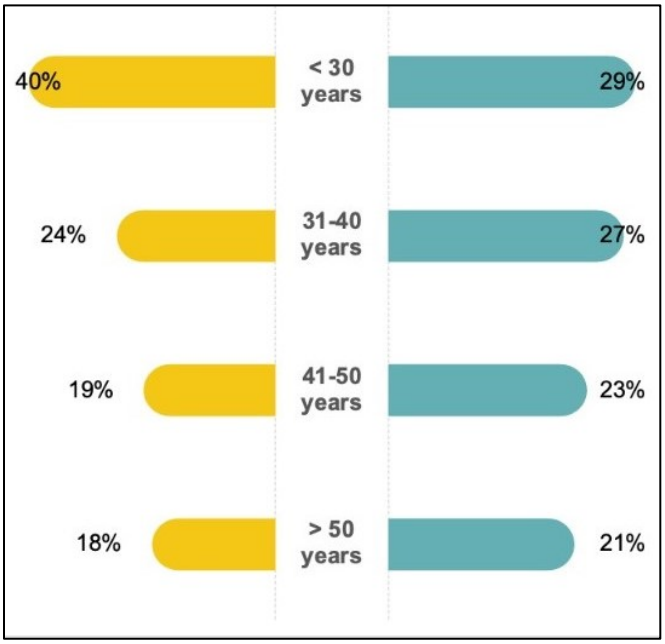
Main dashboard:



Type: Choropleth Map

Encoding: The map illustrates the countries of origin for individuals working in the US who earn an income exceeding $50,000. It is essential to note that our dataset originates from 1994, and for the purposes of this project, countries that no longer exist, such as Yugoslavia, have been excluded. To enhance clarity and analytical capabilities, we have implemented two key filters: (1) Gender-based separation into male and female categories and (2) Geographical separation into continents/regions (Asia, Central America, Europe, North America, South America). Additionally, a feature has been incorporated that allows users to hover on each country, providing detailed information such as the native country's name, the percentage of individuals with an income surpassing $50,000 from that specific country, and the total number of people originating from there. To facilitate user interaction, we are considering the implementation of distinct colors or varying shades of a color for each country, enhancing the ease with which users can click on and distinguish between individual countries.

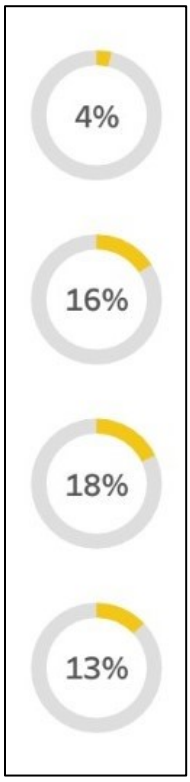
Justification: Geographical data is easier to understand when located in a map. This representation allows the user for quick comprehension of regional income patterns. By incorporating filters for gender and continent/area separation, the map enables multidimensional insights, offering a nuanced understanding of income distribution. Overall, the map serves as a powerful tool for effective communication, informed decision-making, and the identification of trends in income distribution.



Type: Barplot

Encoding: Utilizing horizontal bars, our presentation delineates the percentage distribution of individuals across distinct age intervals (<30 years, 31-40 years, 41-50 years, >50 years). In addition, the gender distribution is visually articulated through the utilization of two specifically chosen colors. It is important to underscore that the scales of the bar plots are dissimilar. In response, we will try to optimize the accuracy of the bar plots to ensure a more precise representation of the respective percentages within the designated age brackets.

Justification: A barplot shows the relationship between a numeric and categorical variable. Even though barplot can considered as a boring chart, is probably the most efficient way to show this kind of data. Therefore, we have determined that a barplot serves as the optimal chart to delineate and illustrate the selected dataset.

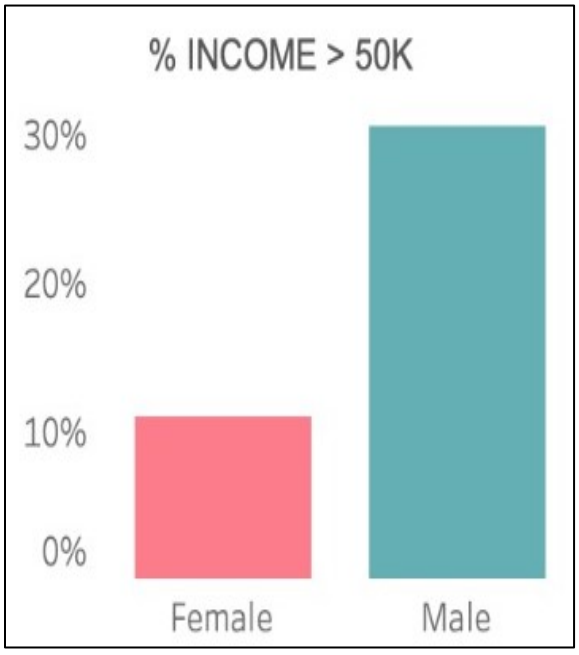


Type: Doughnut chart

Encoding: Employing the doughnut chart, we present a visual representation of the percentage distribution of individuals with an income exceeding $50,000 for each gender. Notably, females are represented by a yellow color, while males are represented by blue. It is noteworthy that these plots are linked to the previously barplot, establishing a comparative framework with each percentage against its corresponding age interval.

Justification: A doughnut plot is employed to visually convey the values of multiple entities, particularly to depict proportions that collectively sum to 100%. This selection is chosen to effectively illustrate the proportional distribution between the two genders within each age group and to elucidate the count of individuals with an income surpassing $50,000.

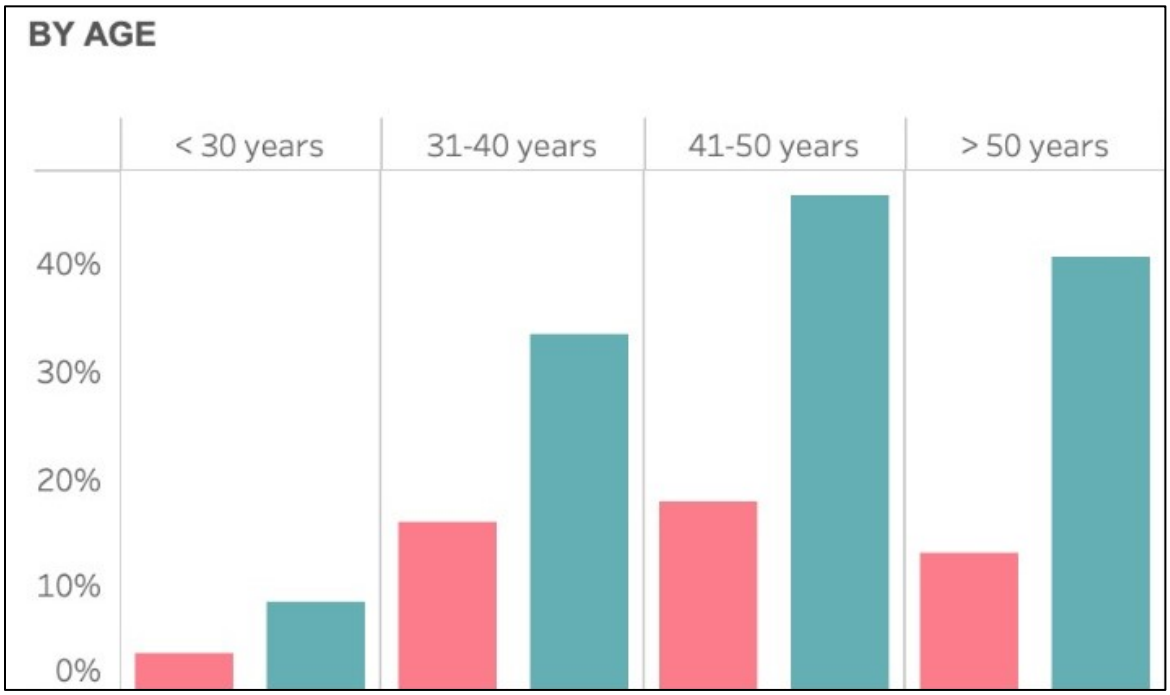
Second Dashboard:



Type: 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%.



Type: Barplot

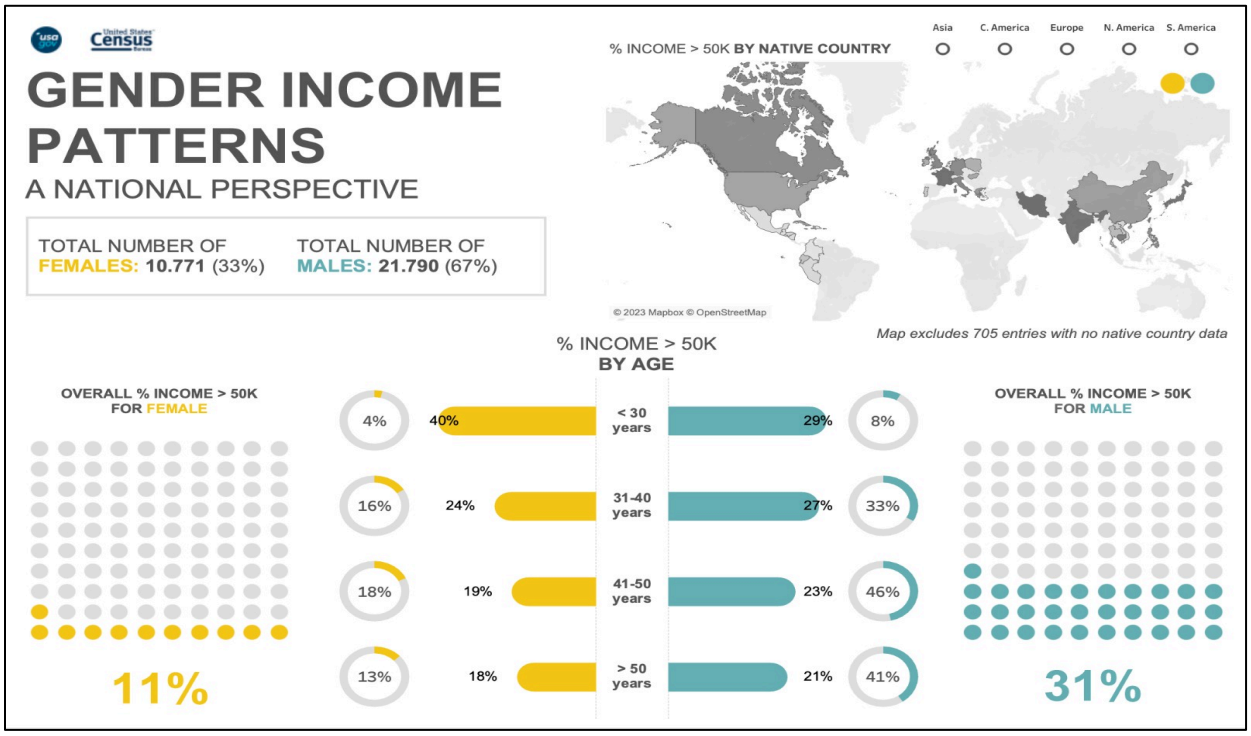
Encoding: In this section, our objective is to distill the same information as presented in the main dashboard, where we segmented the dataset into age intervals (<30 years, 31-40 years, 41-50 years, >50 years) and conducted a comparative analysis between males and females. However, in this segment, 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 maintain a clear and user-friendly interface, we opted for a bar plot in the dashboard. This choice aligns with our goal of providing an easily interpretable and visually accessible solution for users to navigate and extract meaningful insights.

To achieve a high-fidelity design, we have undertaken the refinement of our charts, commencing the planning phase for our ultimate delivery. This involves a consideration of elements crucial to design aesthetics, including scales, labels, colors, and titles, with a particular emphasis on enhancing user perception. As part of this process, we have developed the second iteration of our design concept, ("napkin" design).

Main dashboard:

We removed the filters from the primary dashboard to enhance clarity and provide a comprehensive overview. Subsequently, we propose the incorporation of all filters and detailed information within the secondary dashboard. Additionally, the inclusion of logos representing our client, the U.S. government, and the Census organization, from which our data originates, has been integrated for transparency and acknowledgment of data sources.

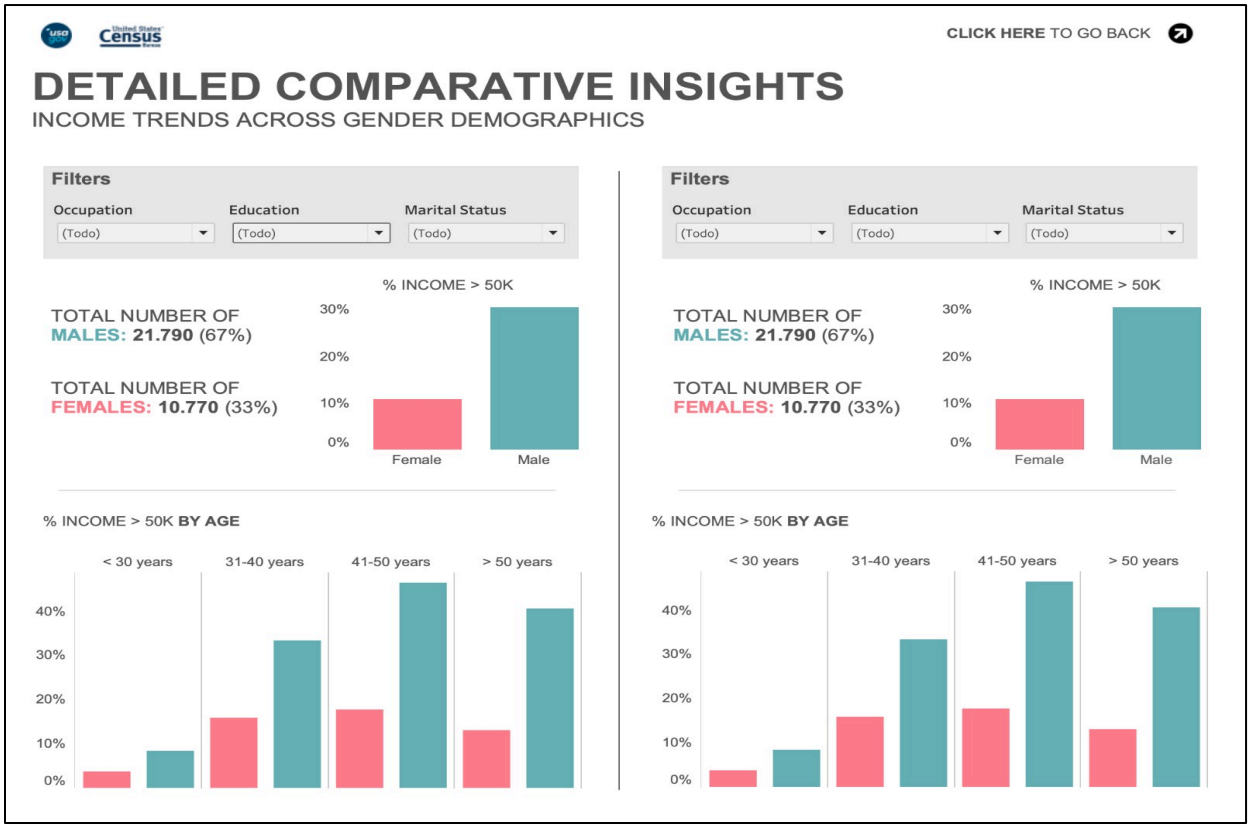


Second dashboard:

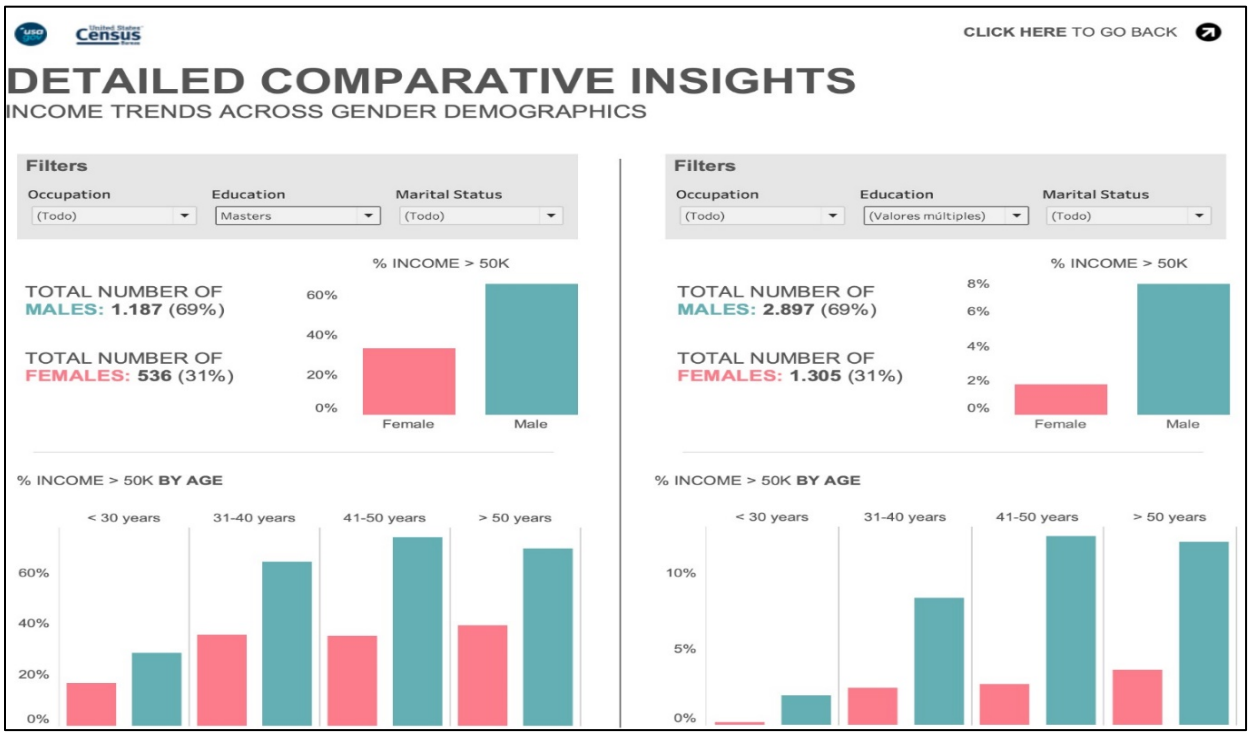
Following a thorough evaluation, we recognized that the initial conception of the second dashboard fell short of anticipated efficiency. Notably, the layout exhibited a level of chaos, impeding the seamless and clear observation of details. Consequently, a collective decision was made to embark on a comprehensive restructuring of the dashboard's framework, while retaining the core idea.

Upon reflection, we decided to split the dashboard into two sections for better comparison among the three key features: education, occupation, and marital status. The original idea was to allow not just cross-feature comparisons but also to delve into specific examples within each feature (we'll get into that later). In essence, it's the same info and ideas as the main dashboard but now neatly organized across all features.

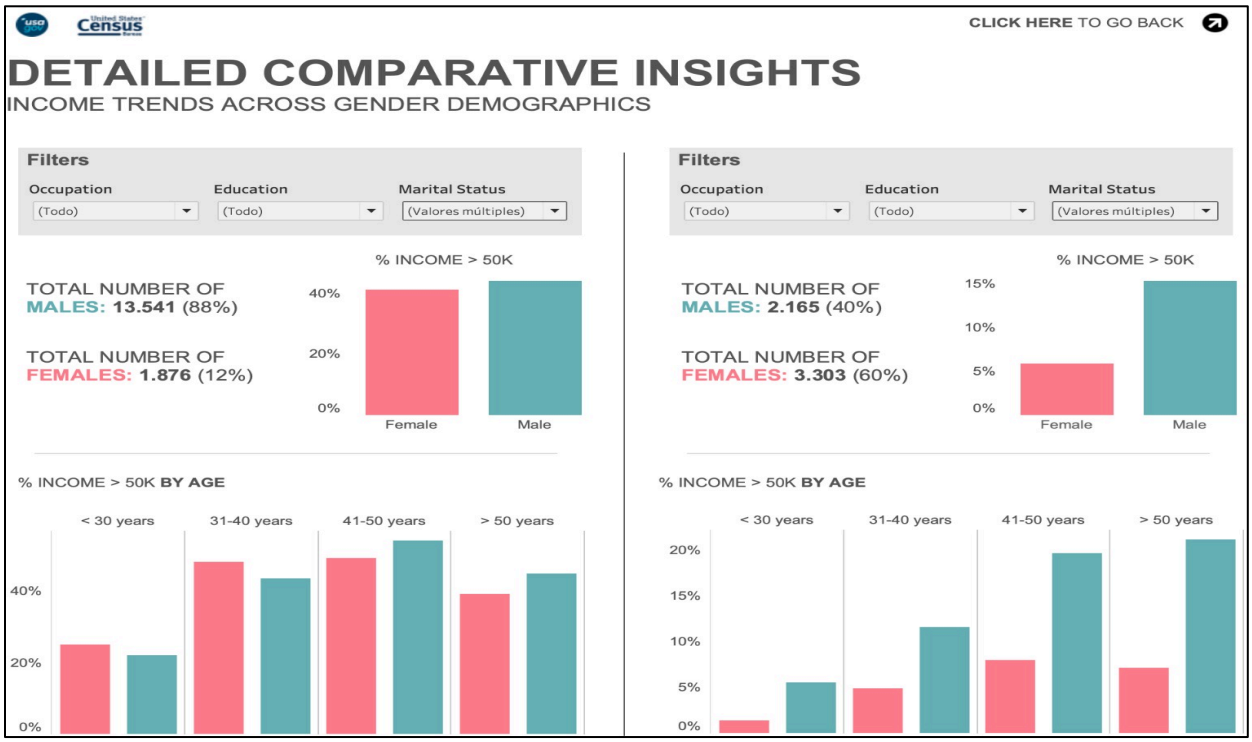
Without applying any filters, the two sides are identical, presenting a setup that looks like the following:



Now, we'll provide some examples to illustrate comparisons within the same features, making it easier to understand:



This comparison focuses on the percentage of people with an income exceeding $50,000, contrasting those with a master's degree and those with education ranging from 1st to 12th grade. Notably, there's a significant shift in the proportion between males and females.



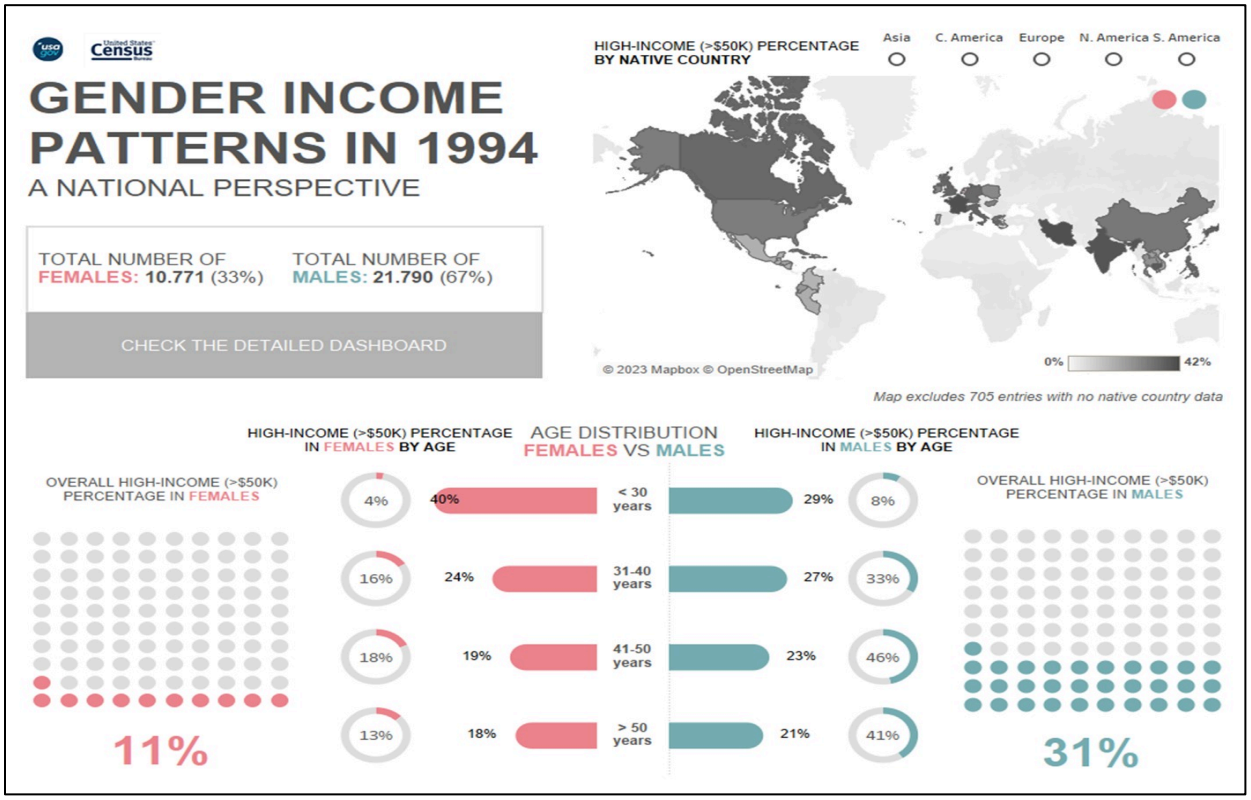
Similarly, when comparing the percentage of individuals with an income surpassing $50,000, we now contrast those who are married with those who are divorced. Once again, a substantial difference is evident in this comparison.

Part 4: Implementation

For the final delivery, our primary emphasis was on refining the suggestions in response to the feedback received from the preceding delivery. Accordingly, we tried to improve our dashboard by implementing various modifications.

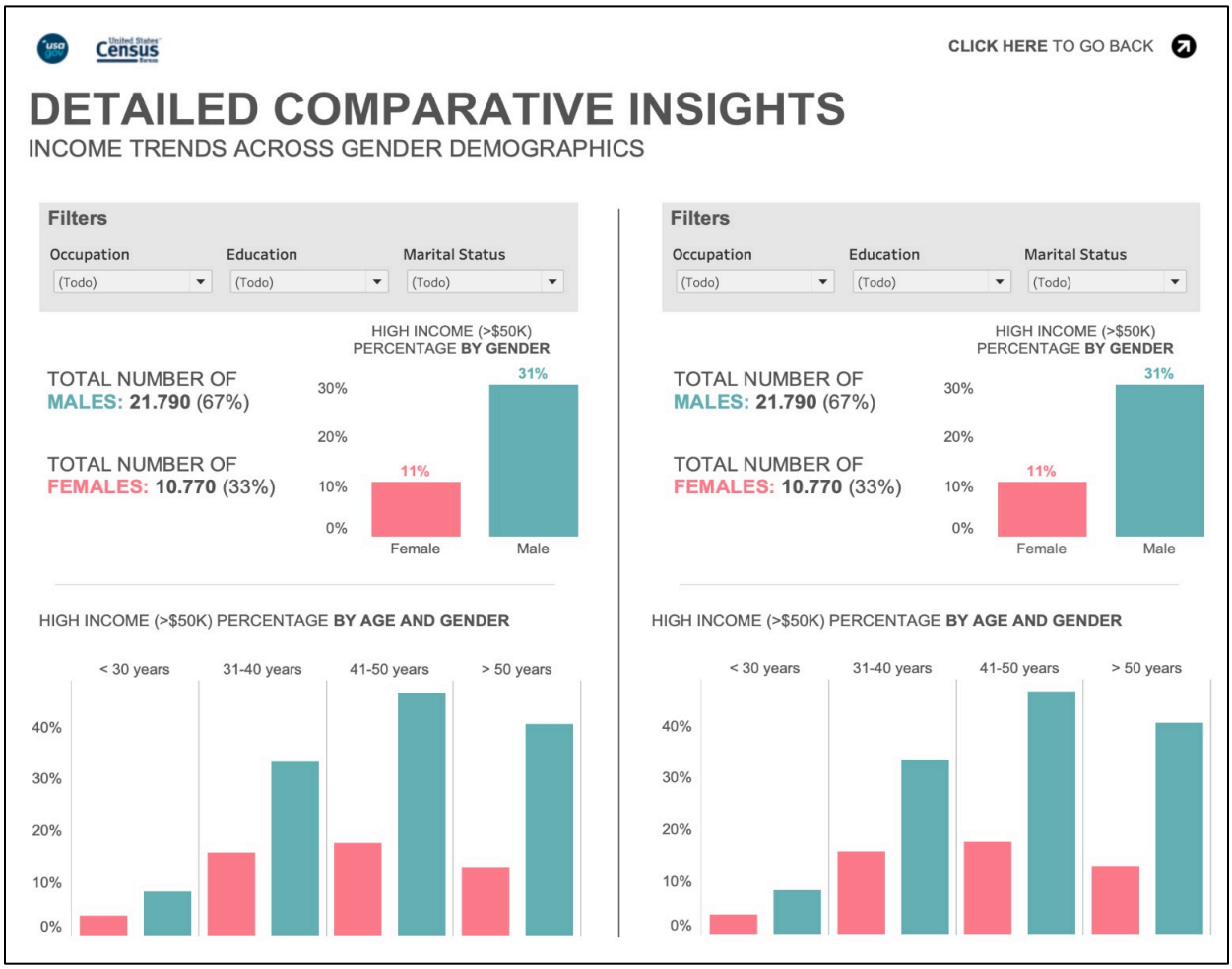
Final main dashboard:

1. We have incorporated links to our dataset and the official website of the Census within the dashboard, facilitating easy access to the data. Users can effortlessly navigate to the respective links by clicking on the associated logos.
2. Enhancements have been made to the clarity and comprehensibility of the dashboard by refining the titles of each chart, along with the main title, ensuring a more precise presentation of information.
3. The colour representation for females has been adjusted to pink, recognizing that the initial choice of yellow may not have been optimal against a white background. This modification was implemented to achieve a color contrast that enhances visibility and legibility.
4. We sorted out the size issues in the bar charts to ensure accurate and consistent representation of the data.
5. We fine-tuned the colour scheme on the map to correspond with the selected gender filter. This means that the map now displays either pink or blue hues based on the chosen gender (male/female). Additionally, the color intensity reflects the percentage of individuals with an income higher $50,000. However, an unresolved issue pertains to the percentage indicator at the bottom of the map, which does not dynamically adjust with the gender selection. This challenge stems from the amalgamation of three distinct graphs layered on top of each other.
6. We introduced a button that allows navigation to the second dashboard.



Final second dashboard:

1. Introduced a navigation button for easy access to the main dashboard.
2. Refined and clarified the titles of each plot for better understanding.
3. Grouped certain data, such as 1st to 12th grade into School, to streamline and improve the overall presentation.
4. Implemented the display of the total number of males and females in the dataset, adjusting the counts based on the applied filters.



Finally, our dashboards empower us to effectively address any lingering inquiries our users may have, offering comprehensive solutions to their previous questions:

- It is worth emphasizing that gender significantly influences income, with females earning considerably less than their male counterparts. Moreover, certain occupations, such as adm-clerical and craft-repair, demonstrate notable income disparities, underscoring the pervasive nature of gender differences across various job categories.
- The distribution of income is evidently influenced by educational attainment, as individuals with higher levels of education tend to earn more. This correlation underscores the impact of education on income levels.
- It appears that marital status plays a significant role in income, with married individuals earning substantially more than their divorced or unmarried counterparts.
- We have discerned that the native country of each individual does not exert any discernible impact on their income.

In conclusion, our designed dashboard exhibits a harmonious structure that prioritizes cleanliness and simplicity across both screens. By emphasizing these qualities, we aim to facilitate seamless user interaction and ensure an effortless journey for extracting the desired conclusions. Through a thoughtful approach to design, we have created an interface that not only provides clarity in data presentation but also enhances the overall user experience, enabling users to effortlessly navigate and derive meaningful insights from the information at hand.