Exploratory Data Analysis and Visualization

02.10.2023
Task 2

Silva, Bruno – UPS10620925
Professor Lac, Visieu

INTERNATIONALE HOCHSCHULE

Table of Contents

List of Figures

Introduction

As per requirement for this assignment, I’ll select a data visualization that, in my assessment, falls short in terms of design and storytelling. I will begin by identifying the shortcomings in the existing visualization. Ultimately, I aim to create a new one that excels in both visualization and storytelling aspects. I'll search for a data visualization on platform Kaggle.

Based on the knowledge I have gained in this course and some insights from self-experience, I will pinpoint and discuss the aspects of the visualization that, in my view, hinder its effectiveness in passing the desired message within the data.

Next, I’ll use the same dataset to create a more powerful and insightful data visualization. I will have in consideration aspects like, color schemes, type of charts and visualization layout, explaining how these contribute to a better data storytelling.

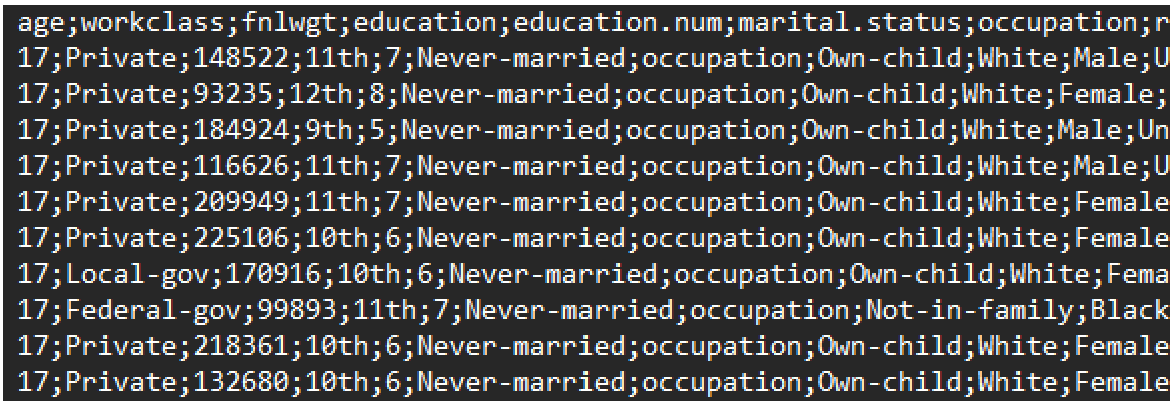
This assignment provides an opportunity for me to refine a data visualization and to deepen my understanding of data science and data storytelling. Upon completion of this assignment, I will have refined my abilities in critical thinking and the enhancement of data visualizations to craft a more compelling narrative.

Kaggle Dataset: https://www.kaggle.com/datasets/anaghakp/adult-income-census/code
Kaggle Visualization: https://www.kaggle.com/code/rajatraj0502/adult-income-census/notebook

1. Sub-Optimal Data Visualization

The first task was to select a sub-optimal data visualization where the dataset was available. As recommended in the assignment I used the Kaggle website to look for this information. After 3 hours and downloading about 20 datasets, I found a Census from 1994 containing the data with people (kaggle, 2023) - https://www.kaggle.com/datasets/anaghakp/adult-income-census/data. As for the data visualization, I selected one where I identified some major problems like, different chart color for the same type of data, overwhelming charts and worst of all, wrong data values.

Starting by the wrong data values, which affects every chart, the selected dataset had a column of weight, meaning, each row repeats itself n number of times where n is the value in this column, look below for a better understanding:



```
age;workclass;fnlwgt;education;education.num;marital.status;occupation;r
17;Private;148522;11th;7;Never-married;occupation;Own-child;White;Male;U
17;Private;93235;12th;8;Never-married;occupation;Own-child;White;Female;
17;Private;184924;9th;5;Never-married;occupation;Own-child;White;Male;Un
17;Private;116626;11th;7;Never-married;occupation;Own-child;White;Male;U
17;Private;209949;11th;7;Never-married;occupation;Own-child;White;Female
17;Private;225106;10th;6;Never-married;occupation;Own-child;White;Female
17;Local-gov;170916;10th;6;Never-married;occupation;Own-child;White;Fema
17;Federal-gov;99893;11th;7;Never-married;occupation;Not-in-family;Black
17;Private;218361;10th;6;Never-married;occupation;Own-child;White;Female
17;Private;132680;10th;6;Never-married;occupation;Own-child;White;Female
```

Figure 1 - Dataset preview

As we can see, the third column ("fnlwgt") has the weight of each row, for example, the first row repats itself 148522 times. In the selected data visualization, the creator ignored this column, creating charts indiscriminately without having an understating of the dataset. In fact, every visualization from different creators (for this dataset) had this mistake. Libraries like matplotlib and seaborn are powerful tools to easily create charts when one row represents one entry in the dataset, but when you have a column for weight, then you must transform your data in order to create accurate charts. The key message I wish to underscore is the importance of not confining your efforts solely to chart creation but rather ensuring a comprehensive understanding of the dataset, thereby guaranteeing the accurate conveyance of information with precise values.

1.1 Age Distribution by Income Group

In our dataset we have a column for the income where it can be “<=50K” or “>50K”, meaning the annual income. Starting by the first chart of our sub-optimal visualization:

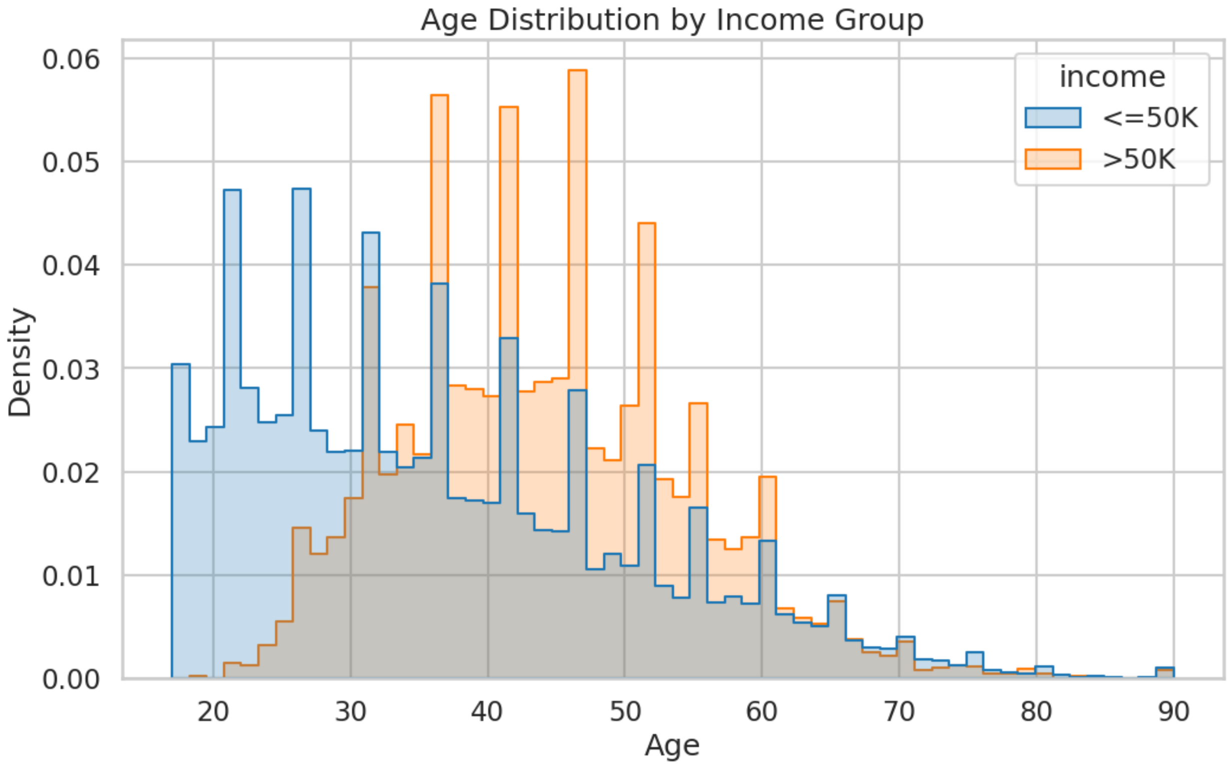


Figure 2 - Age Distribution by Income Group

In this visualization the creator selected 2 colors for each type of income, blue with a darker blue outline for “<=50K” and orange with a darker orange outline for “>50k”. The colors are easy to distinguish between each other, but in the center there is a grey area where both incomes intersect making it harder to see the distribution of our data. I would rather have a bar chart separated by both incomes and have the number of people in the y axis. My produced charts will be on chapter 2.

1.2 Education Level Distribution by Income Group

The second chart is a distribution of the education level separated by the income:

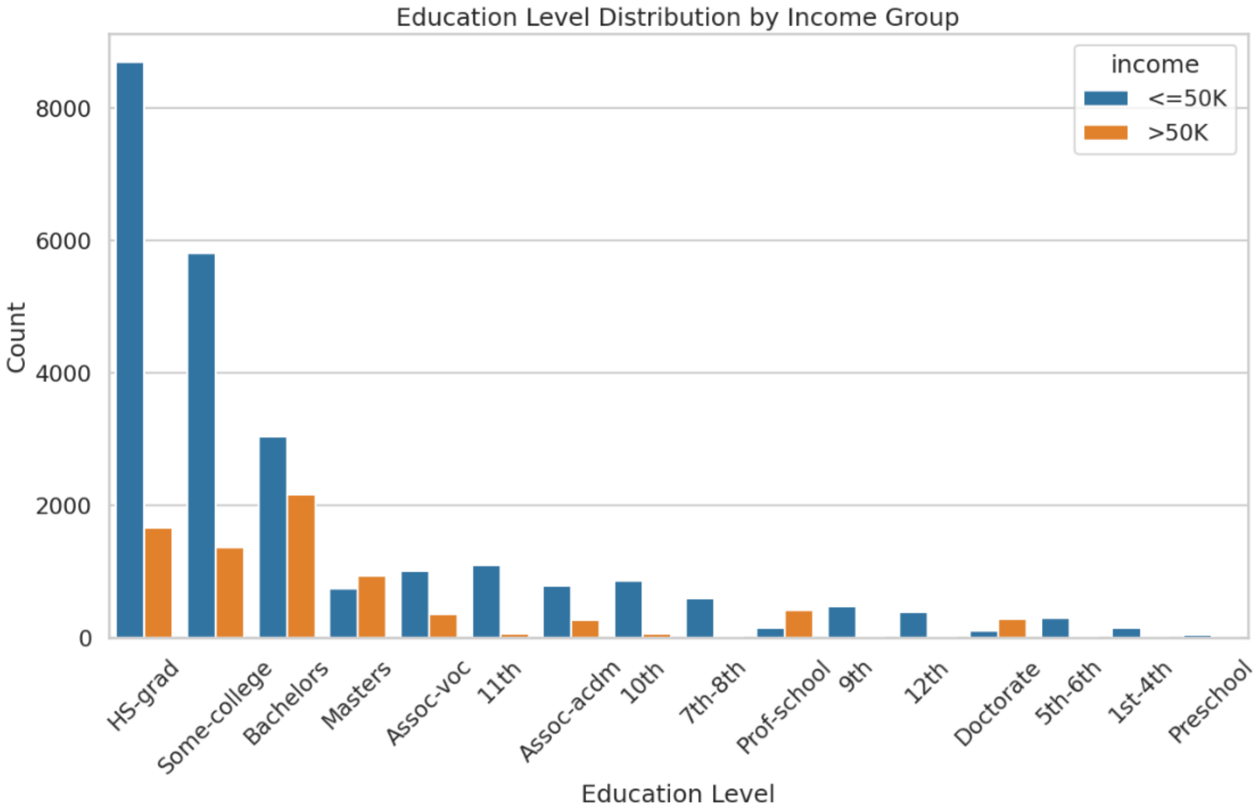


Figure 3 - Education Level Distribution by Income Group

When compared to with the first chart, we see that the colors although the same, they have different transparency and don't have an outline. The chart is ordered by descending "Count" which makes it impossible to see a distribution. The x axis should be ordered by level of education (to represent a distribution) and the y axis should have a better description instead of "Count". I agree with separation of bars for each type of income.

1.3 Workclass Distribution by Income Group

The third chart is a distribution of the workclass separated by the income:

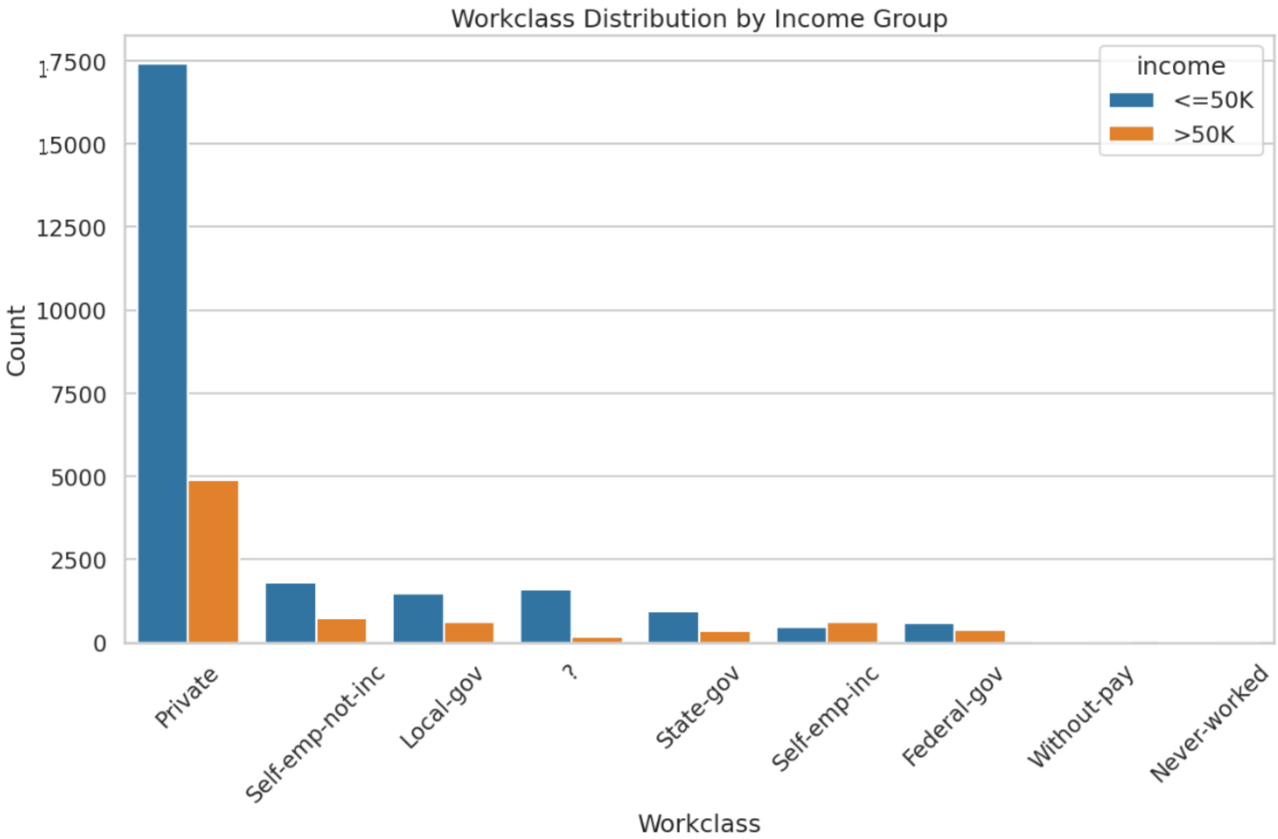


Figure 4 - Workclass Distribution by Income Group

Now we start seeing some consistency in the colors between this chart and the previous one. The main big issue we see here is the “?” on the x axis. We have a sense that most people are in the private sector but when we have a variable called “?” and is bigger than “Without-pay” and “Never- Worked” the logic of the chart falls apart. Once again the y axis description needs to be changed. Because we have the inexplainable variable “?” in the workclass I decided to not use this data at all in my visualization.

1.4 Gender Distribution by Income Group

The fourth chart is a distribution of the gender separated by the income:

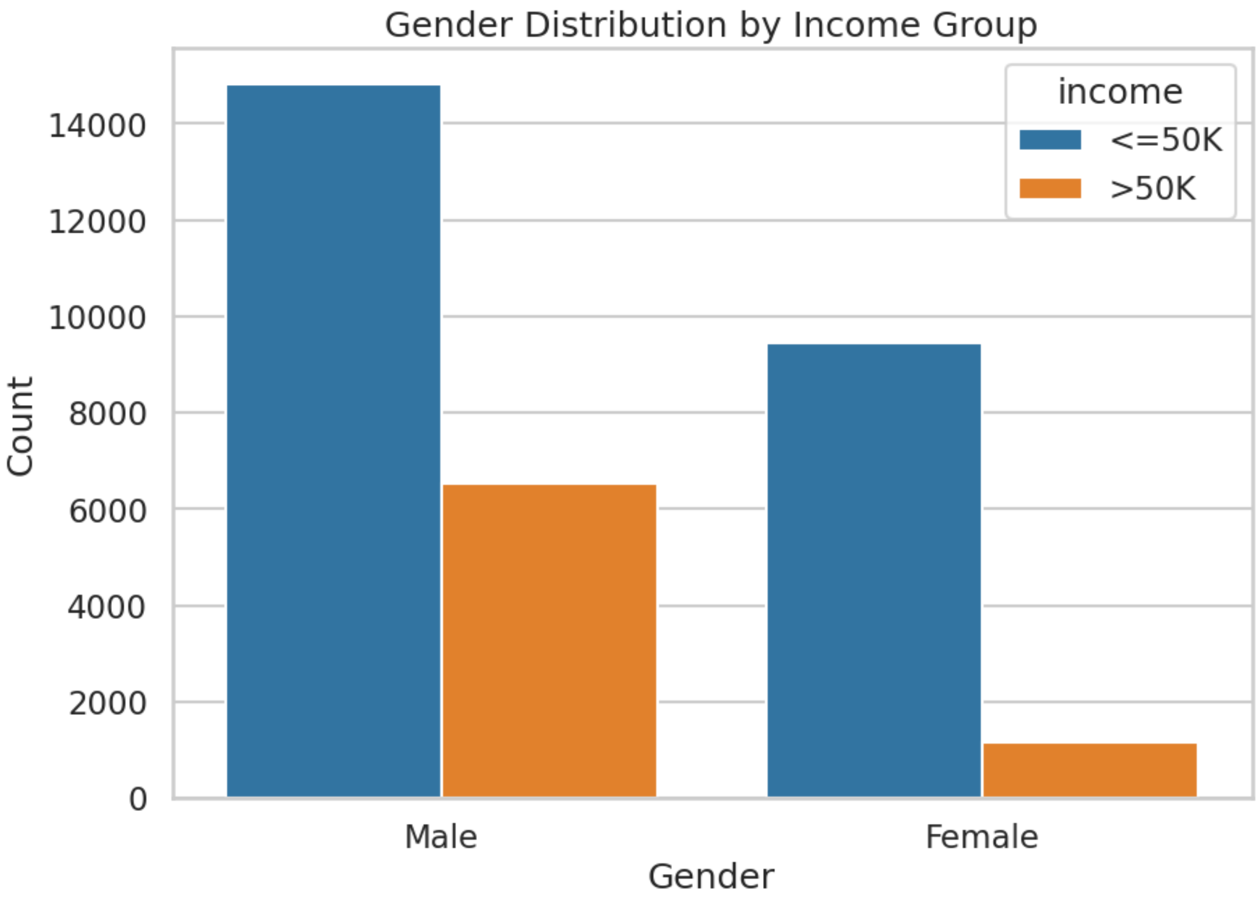


Figure 5 - Gender Distribution by Income Group

This chart is easy to read because it is not overwhelmed with information. A chart presented this way, allows the reader to compare the incomes for each gender and still see that our male population is bigger than female. Once again, the “count” in y axis needs to be changed. If it was not for the wrong numbers here presented (because of the weight), it would be an almost perfect chart. We start to see some consistency between colors on the charts at this point, but soon that is about to change on our sub-optimal visualization.

Because the information from this chart is great to describe our population, I ended up using it, with the correct values.

1.5 Age Distribution by Income Group (2)

The fifth chart is an age distribution by the income group again, this being stacked:

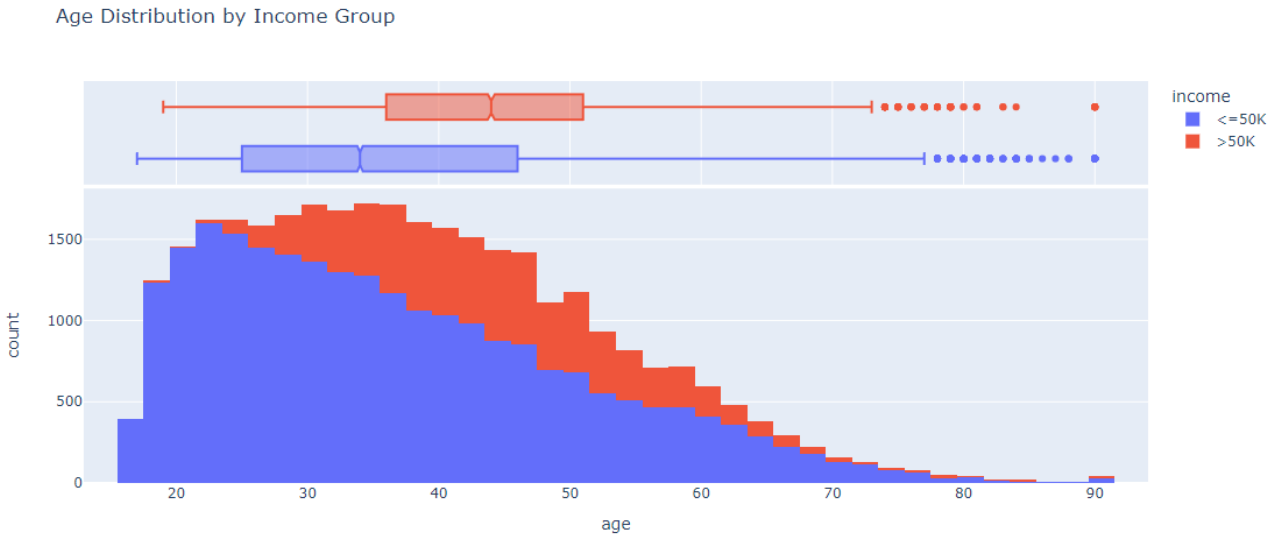


Figure 6 - Age Distribution by Income Group (2)

Going from top to bottom on this figure, we start by seeing a boxplot which I really like, but this type of chart can't be presented to everyone because I requires some kind of knowledge from reader. That being said, it is a good chart, but it was not included in my data visualization. The colors changed from blue and orange to another blue and red. I believe the creator of this visualization used a different library for this chart and just used the default colors of that library. Neglecting the consistency of the colors of our data is going to make our readers lose the focus on our information, instead they are going to be focused in trying to understand the change of colors.

Going to the bottom part of this figure, we have again an age distribution by income group, this time being stacked. I don't really understand how having the same information twice presented in a different way can benefit our story telling. The first time I looked to this chart, it even got me confused because I thought that the red area was behind the blue area, meaning that overall, there were more people earning >50k annually which is not true. I maintain my position, I would rather have a bar chart separated by both incomes and have the number of people in the y axis.

1.6 Education Level Distribution by Income Group (2)

The sixth chart is a distribution of the education level separated by the income again, this time stacked.

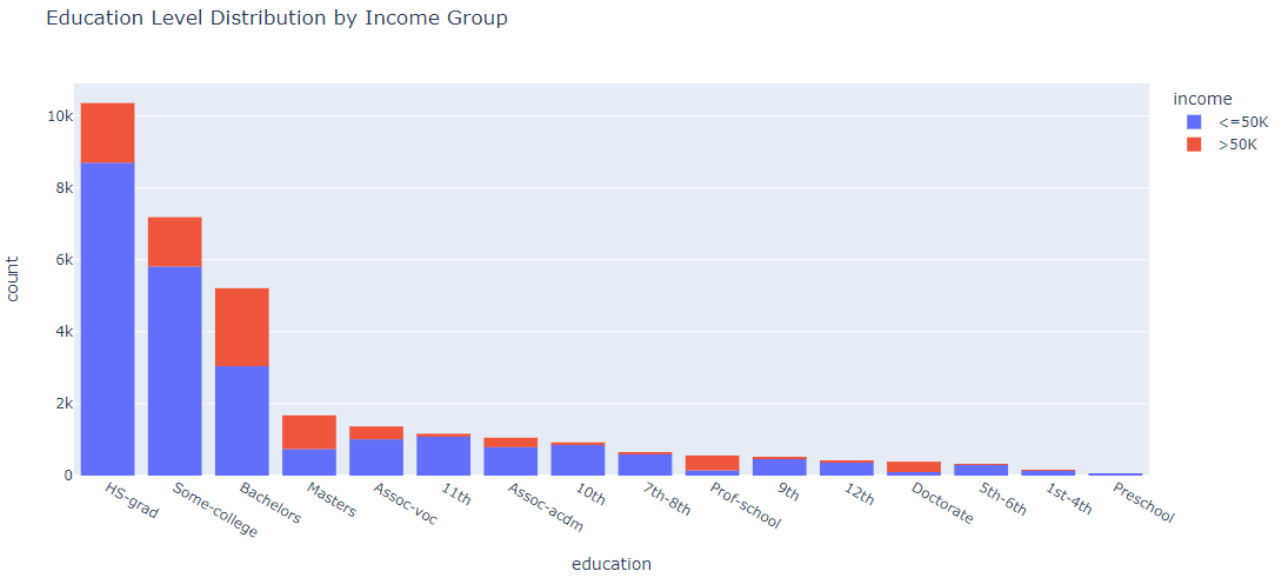


Figure 7 - Education Level Distribution by Income Group (2)

The problem with stacked charts is that it makes it hard to compare, in this case, the population that earn >50K between the different education levels. Once again there is no sense of distribution because the chart is ordered by descending “Count” when it should be ordered by level of education. As said in the previous chart, having the same information represented in a different way, in this case, does not contribute for the progression of our story.

In this sub-optimal visualization, there are multiple charts containing redundant information, which I won't include in this document because it seems repetitive to address the same layout and color corrections.

1.7 Radar Chart Comparing Mean Values by Income Group

The seventh chart is a radar chart to compare means by income:

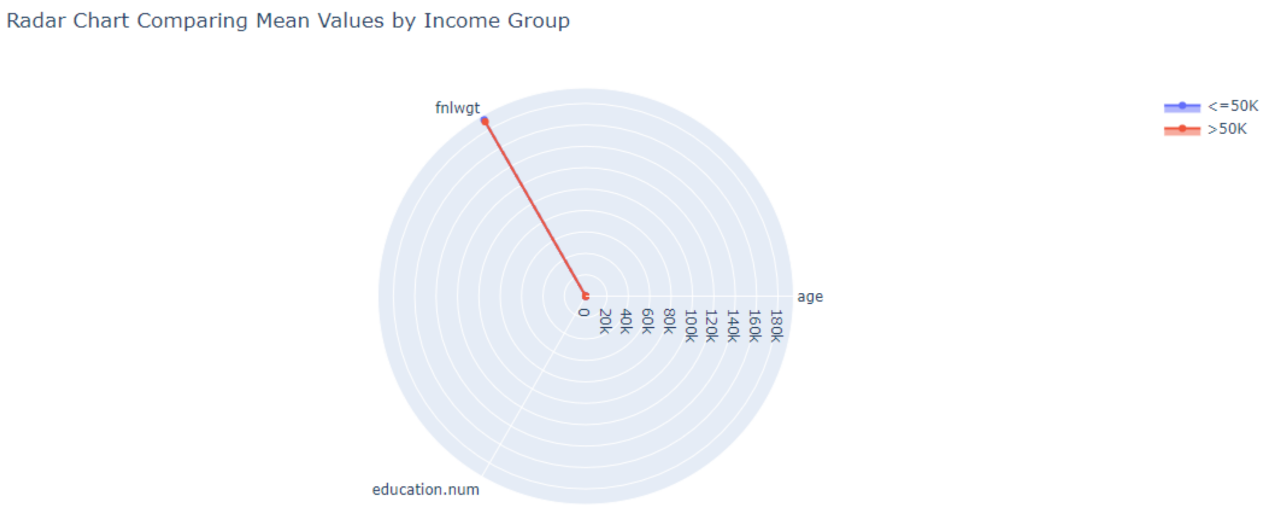


Figure 8 - Radar Chart Comparing Mean Values by Income Group

I can't really understand the purpose of this chart. It looks like for the first time, the variable weight ("fnlwgt") is coming into play but apparently it was badly executed. Looking at the state of this chart, I can't really understand what the creator was trying to do and because of that, I can't really say what could be better. It feels like he was trying to get the age mean per income, education mean per income and fnlwgt per income which makes no sense at all. Once again, the creator was just outputting charts without really understanding the dataset.

1.8 Sunburst Chart: Occupation within Education Level by Income

The eight chart is a sunburst chart with 3 layers, education level, occupation and income:

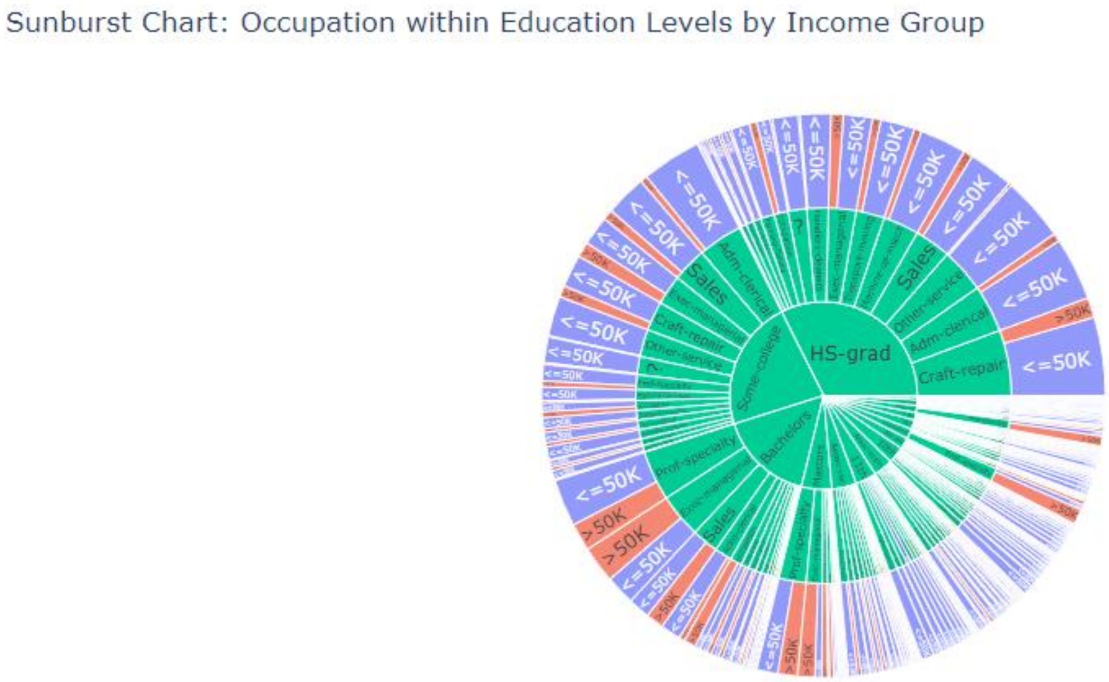


Figure 9 - Sunburst Chart: Occupation within Education Level by Income

The first impression is that we are overwhelmed with information to the point where the reader does not retain anything. If the creator intended to convey a narrative through this chart, I would suggest greying out the entire chart except for the relevant portion that contributes to the story. The main message I see from this chart is that a big part of the population has an HS-graduation and that there is more blue than red, meaning there are more people earning below 50K annually.

1.9 Parallel Categories Diagram of Education, Marital Status, Occupation by Age

The nineth and last chart I will be covering on this document is a parallel chart with the education level, marital status, occupation and income variables by age:

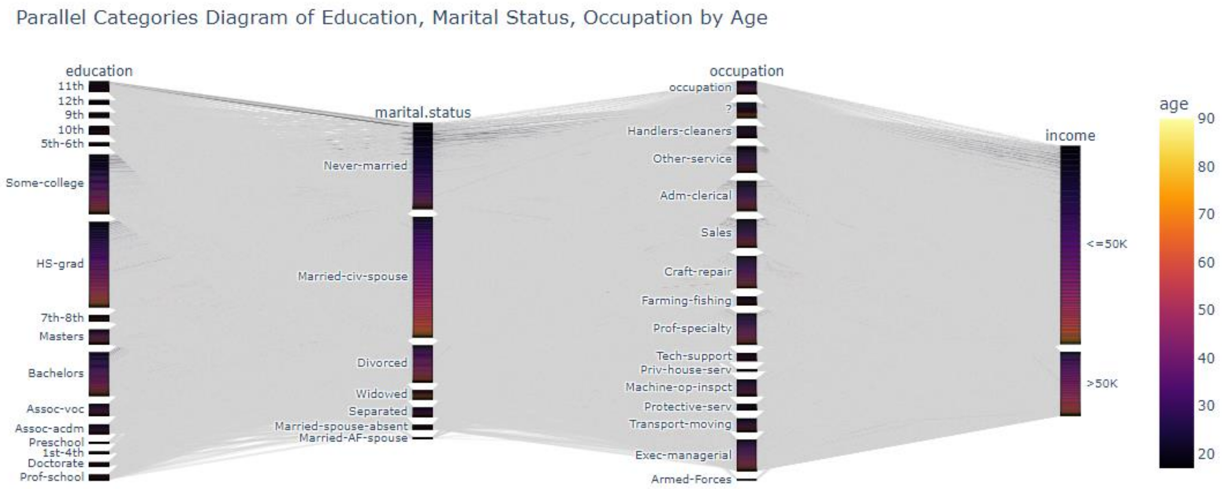


Figure 10 - Parallel Categories Diagram of Education, Marital Status, Occupation by Age

Once more, there is an overwhelming amount of information within the same visualization, making it nearly impossible to discern anything. Most times in data visualization, more ends up being less. The creator should break this chart in multiple smaller visualizations allowing the reader to follow the narrative of the story.

Now that we've examined the issues with this suboptimal visualization, let's address them and work towards a more effective presentation.

2. Improved Data Visualization

This visualization comprises four figures, each containing five charts. We aim to ensure color consistency across the information and maintain a uniform layout, meaning that every figure will adhere to the same layout structure, that being:

1. **Population Description –** We provide an initial overview of our population before delving into deeper analysis. The subsequent sections will focus on an in-depth examination of the top three charts from this overview.
2. **Gender Analysis –** Selecting the first chart from the previous figure, we give the reader a deeper analysis between the genders.
3. **Income Analysis** - Selecting the second chart from the first figure, we give the reader a deeper analysis between the two types of income.

4. **Ethnicity Analysis** - Selecting the third chart from the first figure, we give the reader a deeper analysis between the different ethnicities.

Additionally, it's worth noting that the forthcoming figures will take into account the weight column, ensuring we provide the accurate information to the reader.

2.1 Population Description

The first figure, Population Description:

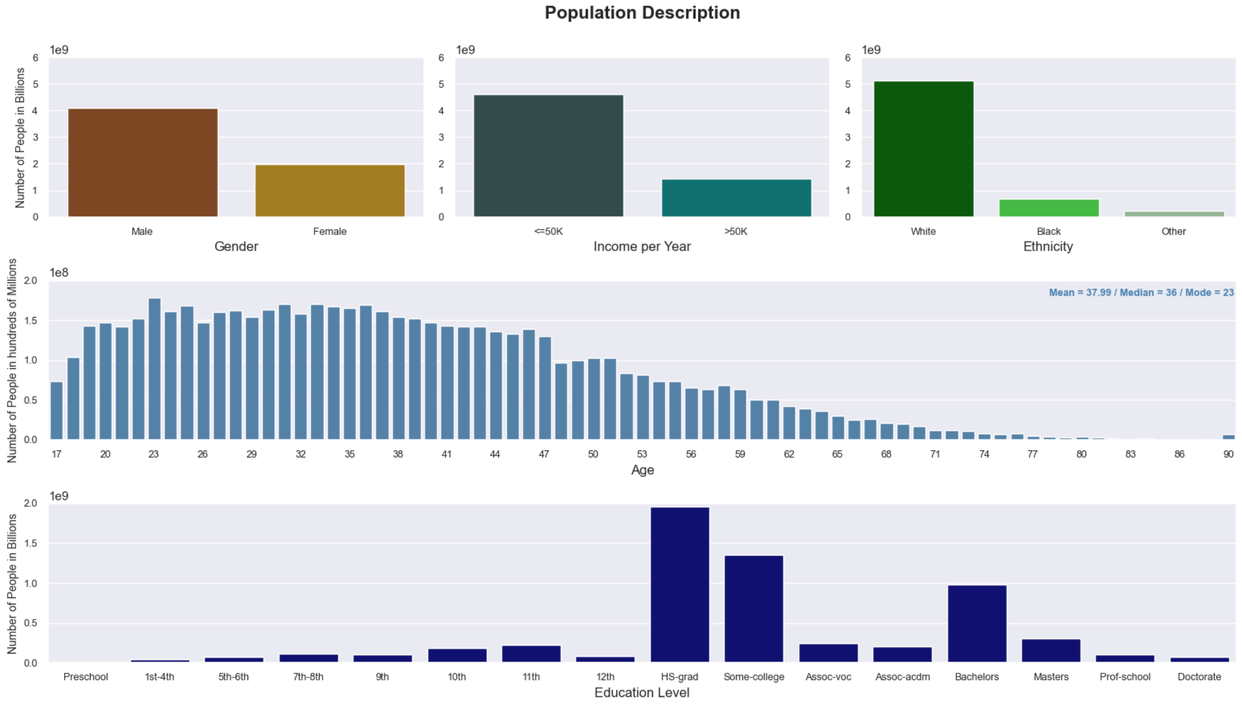


Figure 11 - Population Description

Dividing our figure into five charts:

1. **Gender Chart:** The y axis represents the population in billions and that description extends horizontally to the charts on the right. We have a brown color for male and dark yellow for female. Notably, for this particular population, we observe that the number of males is more than twice that of males.
2. **Income per Year Chart:** For the income “<=50K” we have the color “darkslategrey” and for “>50K” we have the color “teal”. The number of people earning more than 50K annually is 3 times smaller than the number of people earning less than 50K.
3. **Ethnicity Chart:** In this chart, while there were several distinct ethnicities, due to their small individual quantities, they have been combined into a single "Other" category on the bar graph. This approach minimizes the clutter of additional bars, ensuring that

the reader can maintain focus on the key aspects we wish to convey. The population is predominately "white", with "black" being the second largest group.

4. **Age Chart:** This time the y axis represents the population in hundreds of millions. We have expanded the chart horizontally to accommodate the numerous age groups on the x-axis, enhancing readability. It's evident that our population exhibits a left-skewed distribution, signifying higher density among younger ages. We also have the mean, median and mode on the top-right side of the chart which had to be calculated without using libraries because of the weight column.
5. **Education Level Chart:** The x axis is ordered by the education level. I didn't define this order, our dataset had two columns, one with the education level number (between 1 and 16) and one with the education level name. The main idea the reader takes from this chart is that the 3 bigger groups are "HS-grad", "Some-college" and "Bachelors". If you remember, our sub-optimal visualization was ordered by the number of people, making it impossible to differentiate between education levels.

2.2 Gender Analysis

The second figure, a deeper Gender Analysis:

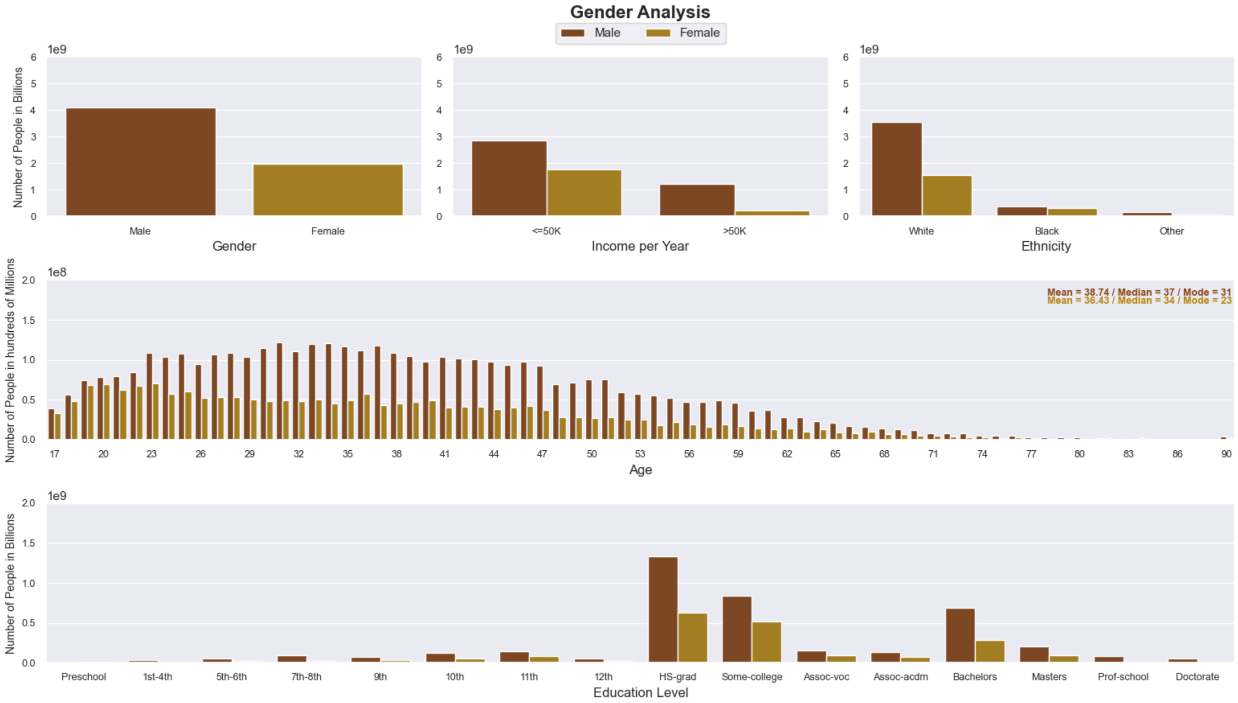


Figure 12 - Gender Analysis

As we can see, this figure respects the same layout and colors as the previous figure (brown for male.and dark yellow for female) Dividing our figure into the five charts:

1. **Gender Chart:** It is exactly the same chart from the previous figure, but it is important to include it in this visualization because this figure is all about Gender analysis.
2. **Income per year per Gender Chart:** The key takeaway from this chart is the stark gender disparity in earnings, particularly evident in the fact that a significantly smaller number of females earn above 50K. Additionally, it highlights a substantial income gap between males and females, with the income distribution for males showing much less variation compared to that of females.
3. **Ethnicity per Gender Chart:** Among individuals of "white" ethnicity, around 30% are female, whereas within the "black" ethnicity, the gender distribution is notably more balanced, with a smaller gap between males and females.
4. **Age per Gender Chart:** Both distributions exhibit left-skewness, with females having a relatively younger distribution compared to males. The female distribution appears more linear, while the male distribution exhibits a subtle curve. It's important to bear in mind that this dataset dates back to 1994 and exclusively comprises individuals with an annual income (AGI) surpassing $100. This likely explains the elevated figures observed in younger females and keeps dropping as age goes up. This pattern can be related to historical societal roles in certain countries where women were primarily viewed as housekeepers.
5. **Education Level by Gender Chart:** There are no prominent highlights to extract from this chart, except for the similarity on the distribution of both females and males, despite the fact that there are less females.

2.3 Income Analysis

The third figure, a deeper Income Analysis:

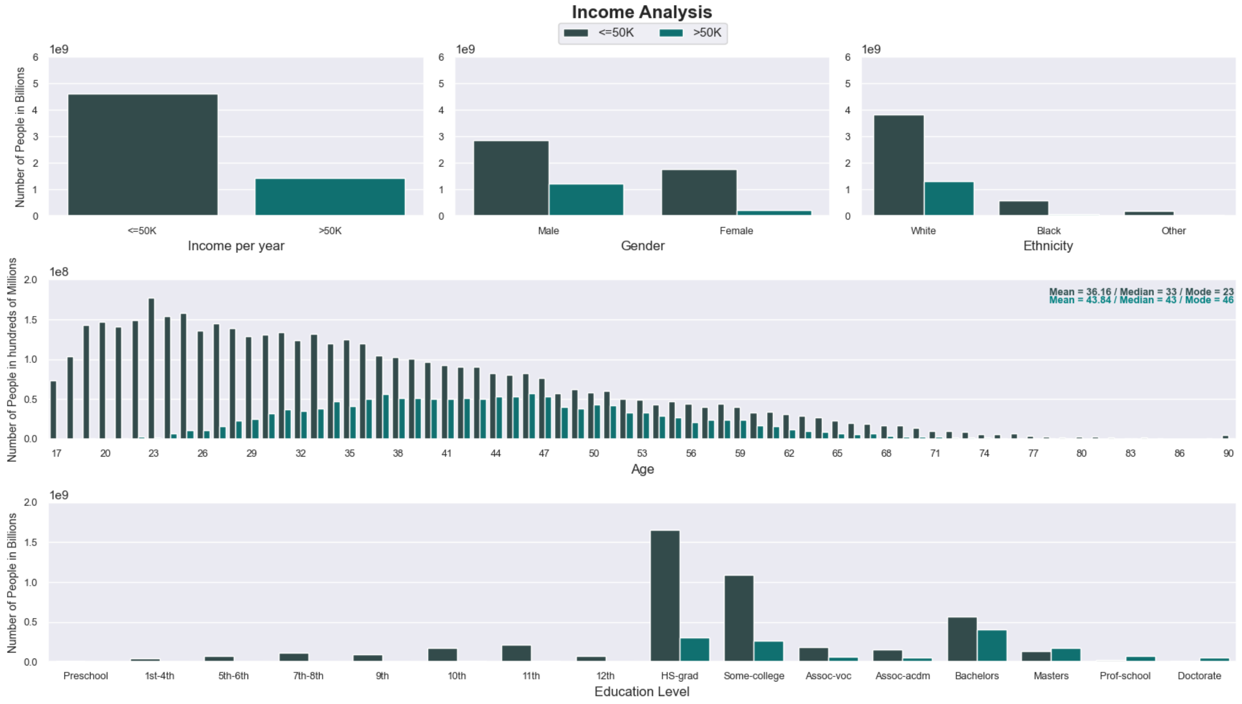


Figure 13 - Income Analysis

Once again, same layout and same colors for each type of income. From all the figures, I think these two colors are the ones that combined the best together.

1. **Income per Year Chart:** Again, the first chart comes from the first figure.
2. **Geder per Income Chart:** It is evident that the disparity in earnings between females is more pronounced compared to that observed among males.
3. **Ethnicity per Income Chart:** Among all ethnicities, the "white" population stands out with the largest portion of individuals earning more than 50K.
4. **Age per Income Chart:** In the case of earnings categorized as "<=50K," we observe a left-skewed distribution, indicating its prevalence among younger generations. Conversely, for earners with ">50K," the distribution is closer to normal, suggesting a higher prevalence among individuals around the age of 44, in the middle of their careers.
5. **Education Level per Income Chart:** we can observe that for the last three levels of education (Master's, Prof-school, and Doctorate), there is a higher number of individuals earning ">50K" compared to the other income category, which indicates that the proportion of individuals earning ">50K" tends to increase as the level of education rises.

2.4 Ethnicity Analysis

The fourth figure, a deeper Ethnicity Analysis:

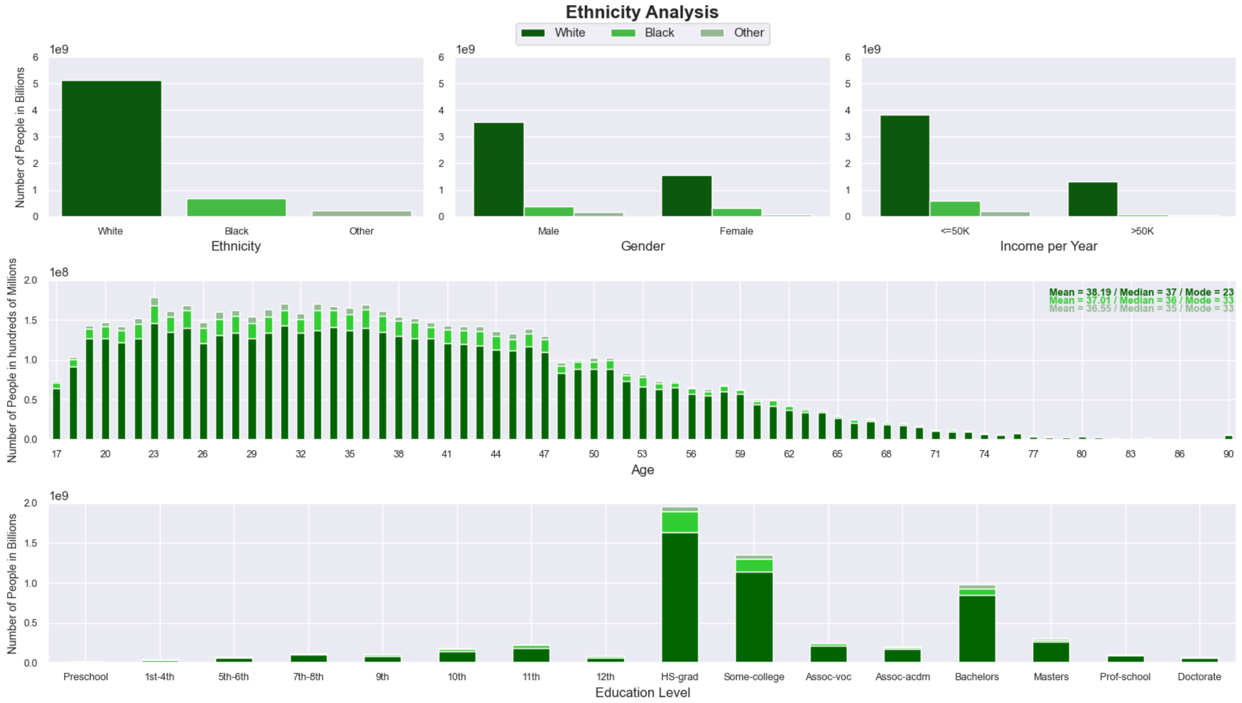


Figure 14 - Ethnicity Analysis

1. **Ethnicity Chart:** Again, the first chart comes from the first figure.
2. **Gender per Ethnicity Chart:** Across all ethnicities, there is a consistent balance in the proportions between male and female groups.
3. **Income per Ethnicity Chart:** The ethnicities “Black” and “Other” are almost inexistent in the group that earn above 50K. The proportions for each ethnicity are very different for each income group.
4. **Age per Ethnicity Chart:** I opted to use stacked bars for this chart because presenting individual bars for each category would have made it too crowded and challenging to interpret. However, by using stacked bars, we sacrifice the ability to visualize the specific distribution for each ethnicity. Once again, we observe a left-skewed distribution, indicating a higher concentration in younger generations. Notably, among the "other" ethnicities, the distribution is relatively more uniform across age groups until it begins to decline around the age of 50, mirroring the pattern seen in other ethnicities.
5. **Education Level per Ethnicity Chart:** As mentioned in the previous figure, it was noted that individuals with the last three levels of education were more likely to have incomes above 50K. In this chart, we can observe that the majority of people with these higher education levels belong to the "white" ethnicity, meaning that this group has higher levels of education and higher incomes.

Conclusion

In this project, I initiated by identifying and critiquing key flaws in a sub-optimal visualization, which subsequently led me to create my own improved version. The major errors we observed in the sub-optimal visualization included inconsistent color usage for the same variables, improperly ordered charts, excessive information making some charts unreadable, and a significant issue where the data presented to the reader was inaccurate due to the creator's oversight of the weight column.

Once I had identified and structured the main issues, it became easier to create an improved visualization. I began by organizing the relevant information for each figure, which involved a lengthy process. I experimented with various color combinations and different figure layouts along the way, ultimately leading to this final version of the visualization.

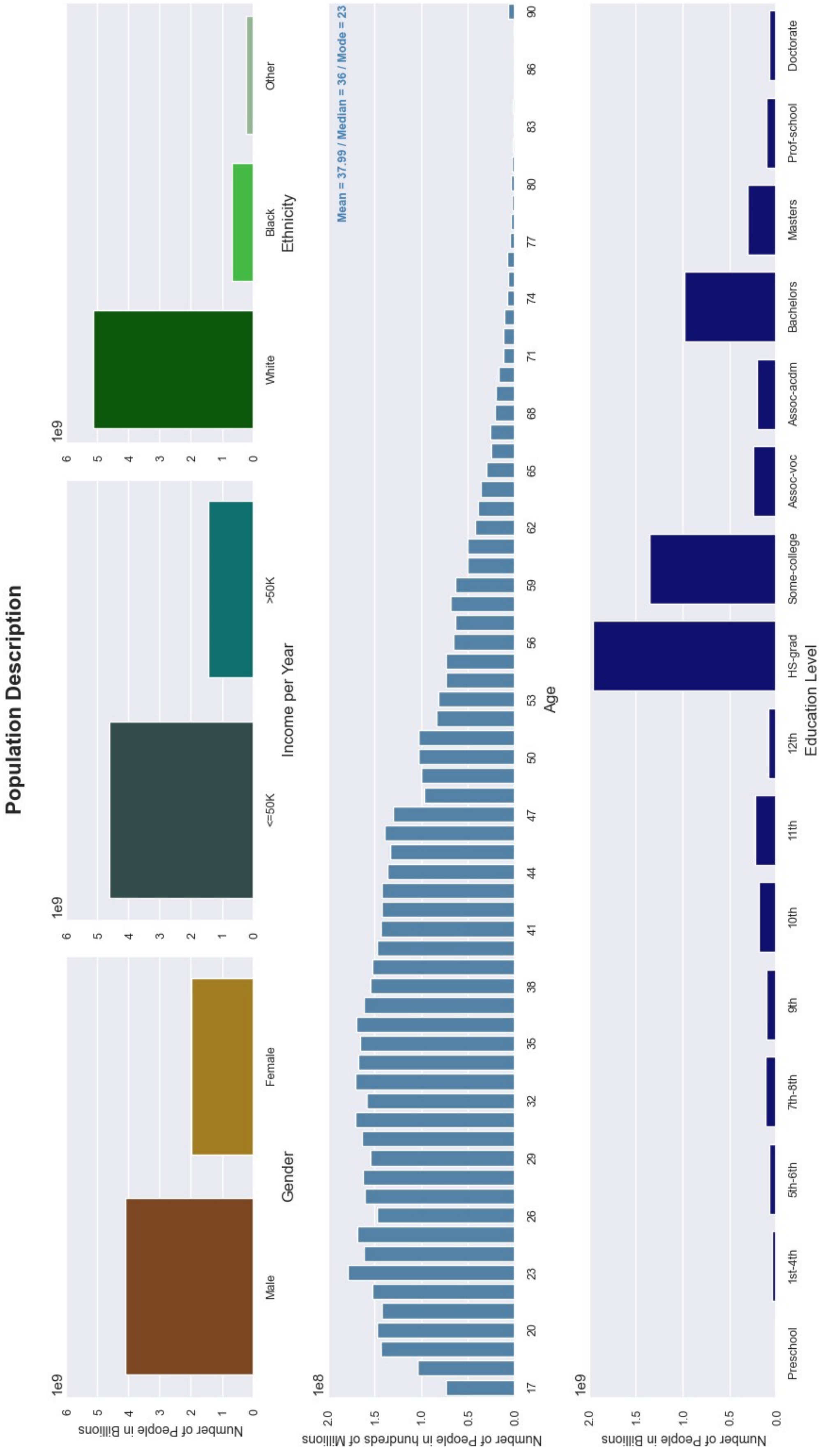
Wanted to mention as well that my calculations of mean, median and mode were done manually without the use of libraries because the dataset was not properly set for that end (weight column).

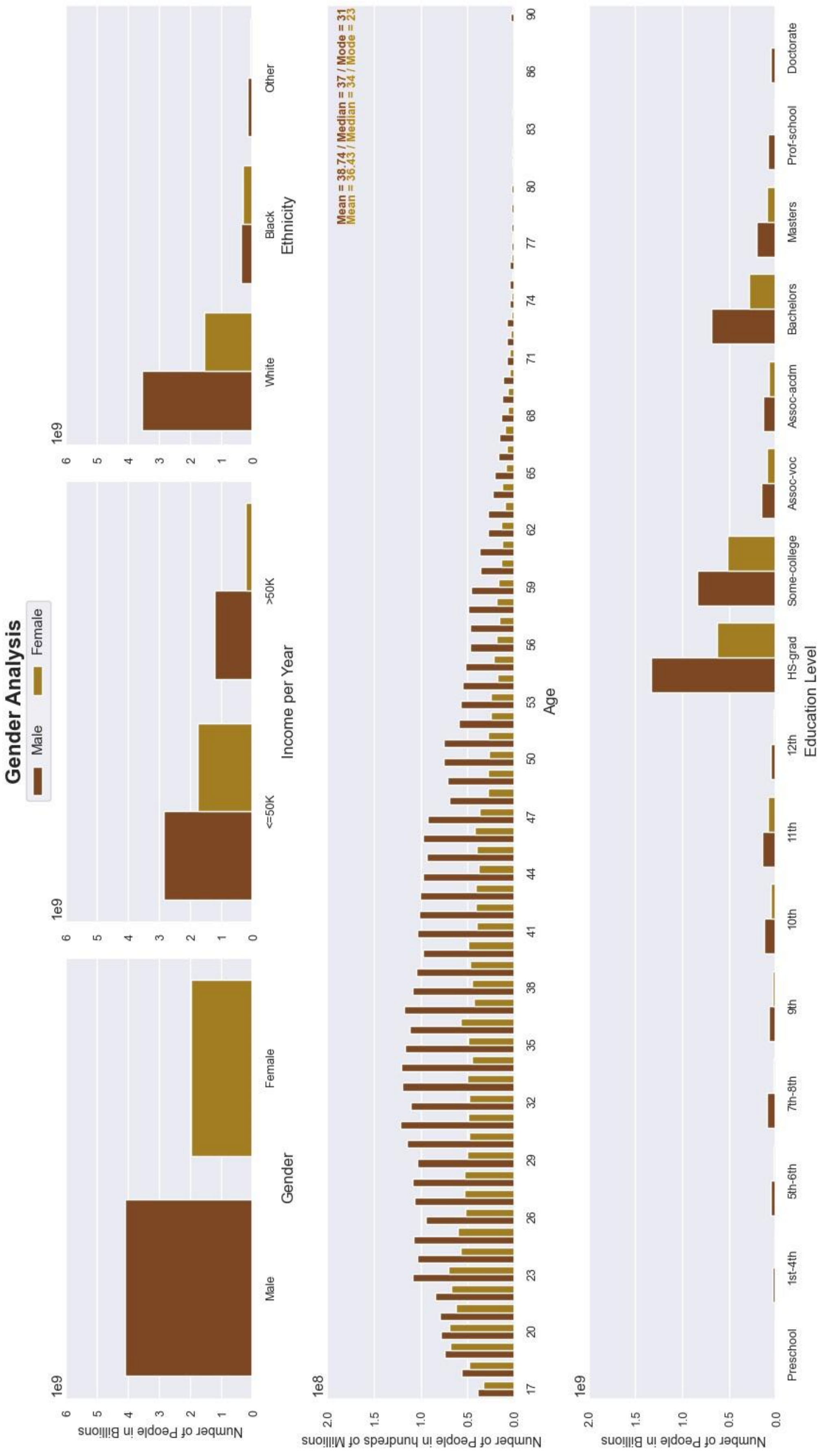
In the end, I am satisfied with the final result of the visualization. However, through experience, I've learned that a recurring visualization often undergoes numerous iterations over time before achieving the perfect balance of aesthetics and information delivery.

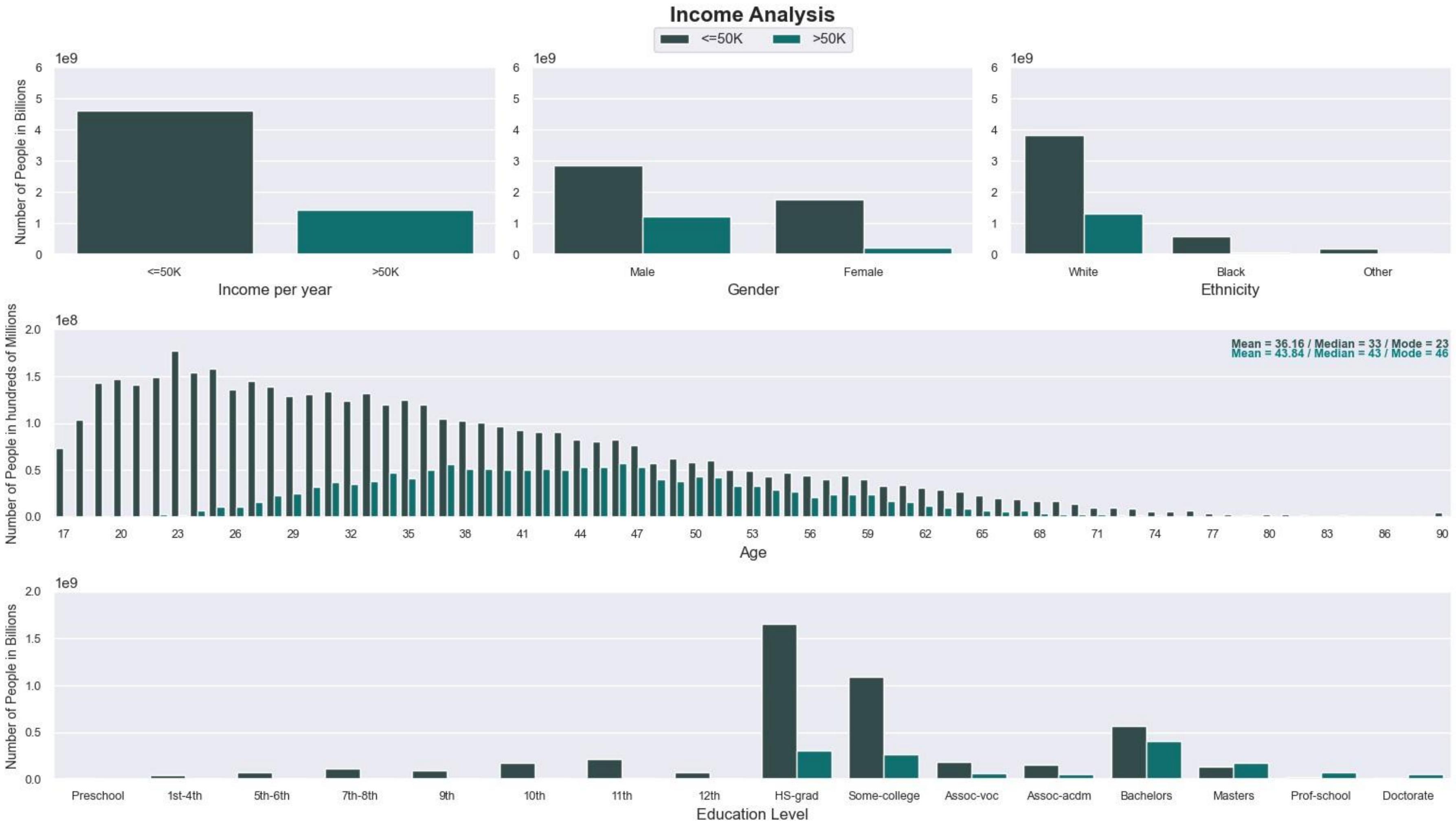
Bibliography

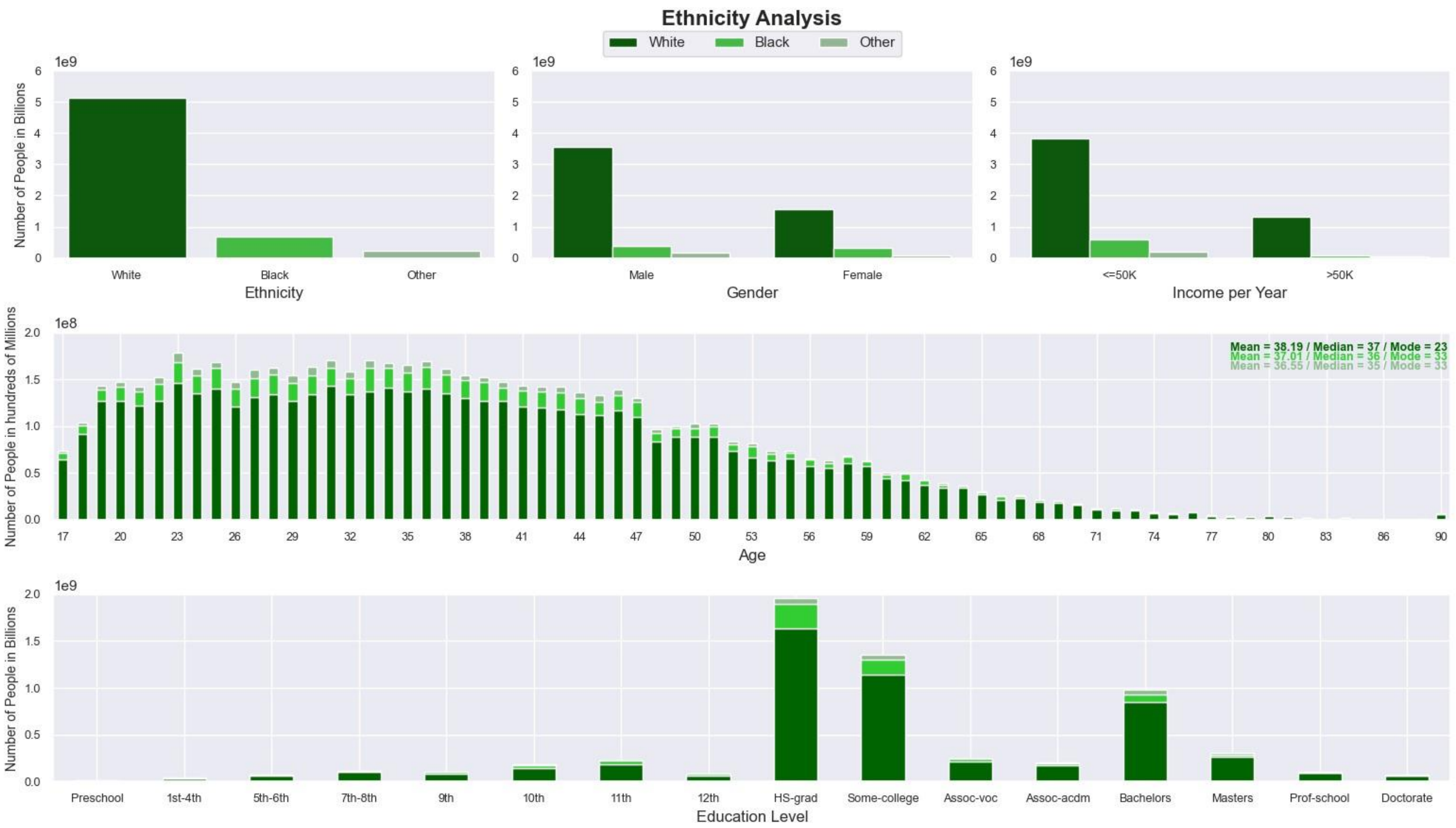
kaggle. (2023, September 7). Retrieved from kaggle: https://www.kaggle.com/datasets/anaghakp/adult-income-census/data

Appendices


Population Description
Number of People in Billions
1e9
6
5
4
3
2
1
0
Male
Female
Gender
1e9
Income per Year
<=50K
>50K
1e9
Ethnicity
White
Black
Other
Number of People in hundreds of Millions
1e8
2.0
1.5
1.0
0.5
0.0
Mean = 37.99 / Median = 36 / Mode = 23
Age
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71
74
77
80
83
86
90
Number of People in Billions
1e9
2.0
1.5
1.0
0.5
0.0
Education Level
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS-grad
Some-college
Assoc-voc
Assoc-acdm
Bachelors
Masters
Prof-school
Doctorate


Gender Analysis
Male
Female
1e9
Number of People in Billions
0
1
2
3
4
5
6
Male
Female
Gender
1e9
<=50K
>50K
Income per Year
1e9
White
Black
Other
Ethnicity
1e8
Number of People in hundreds of Millions
0.0
0.5
1.0
1.5
2.0
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71
74
77
80
83
86
90
Age
Mean = 38.74 / Median = 37 / Mode = 31
Mean = 36.43 / Median = 34 / Mode = 23
1e9
Number of People in Billions
0.0
0.5
1.0
1.5
2.0
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS-grad
Some-college
Assoc-voc
Assoc-acdm
Bachelors
Masters
Prof-school
Doctorate
Education Level


Income Analysis
<=50K
>50K
1e9
0
1
2
3
4
5
6
Number of People in Billions
<=50K
>50K
Income per year
Male
Female
Gender
White
Black
Other
Ethnicity
1e8
Number of People in hundreds of Millions
0.0
0.5
1.0
1.5
2.0
Mean = 36.16 / Median = 33 / Mode = 23
Mean = 43.84 / Median = 43 / Mode = 46
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71
74
77
80
83
86
90
Age
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS-grad
Some-college
Assoc-voc
Assoc-acdm
Bachelors
Masters
Prof-school
Doctorate
Education Level


Ethnicity Analysis
White
Black
Other
1e9
Number of People in Billions
0
1
2
3
4
5
6
White
Black
Other
Ethnicity
1e9
Male
Female
Gender
1e9
<=50K
>50K
Income per Year
1e8
Number of People in hundreds of Millions
0.0
0.5
1.0
1.5
2.0
Mean = 38.19 / Median = 37 / Mode = 23
Mean = 37.01 / Median = 36 / Mode = 33
Mean = 36.55 / Median = 35 / Mode = 33
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71
74
77
80
83
86
90
Age
1e9
Number of People in Billions
Preschool
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
HS-grad
Some-college
Assoc-voc
Assoc-acdm
Bachelors
Masters
Prof-school
Doctorate
Education Level