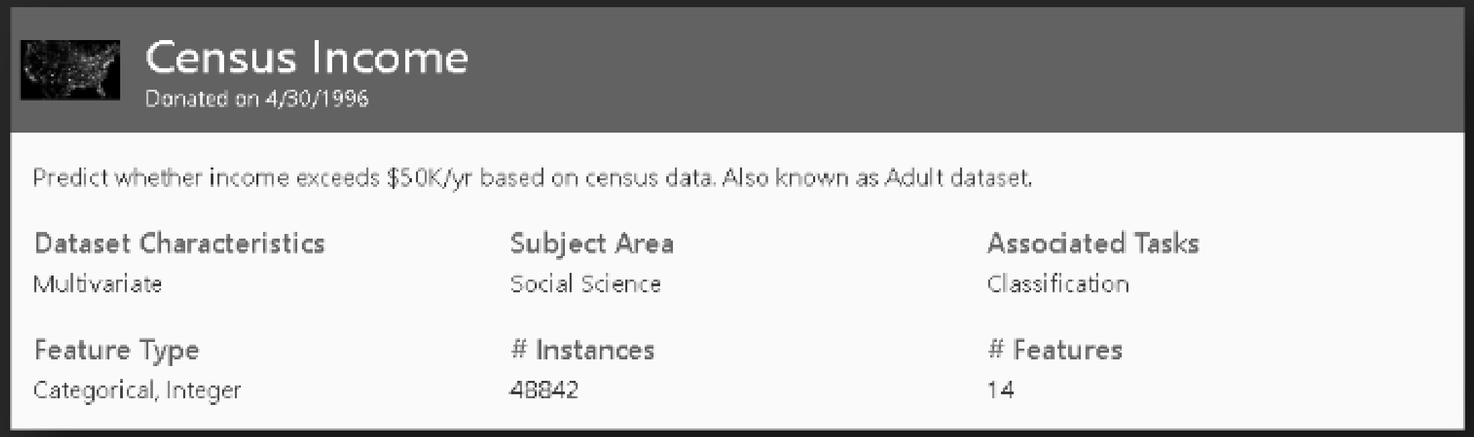
Midterm Skills Exam: Data Wrangling and Analysis

CENSUS INCOME

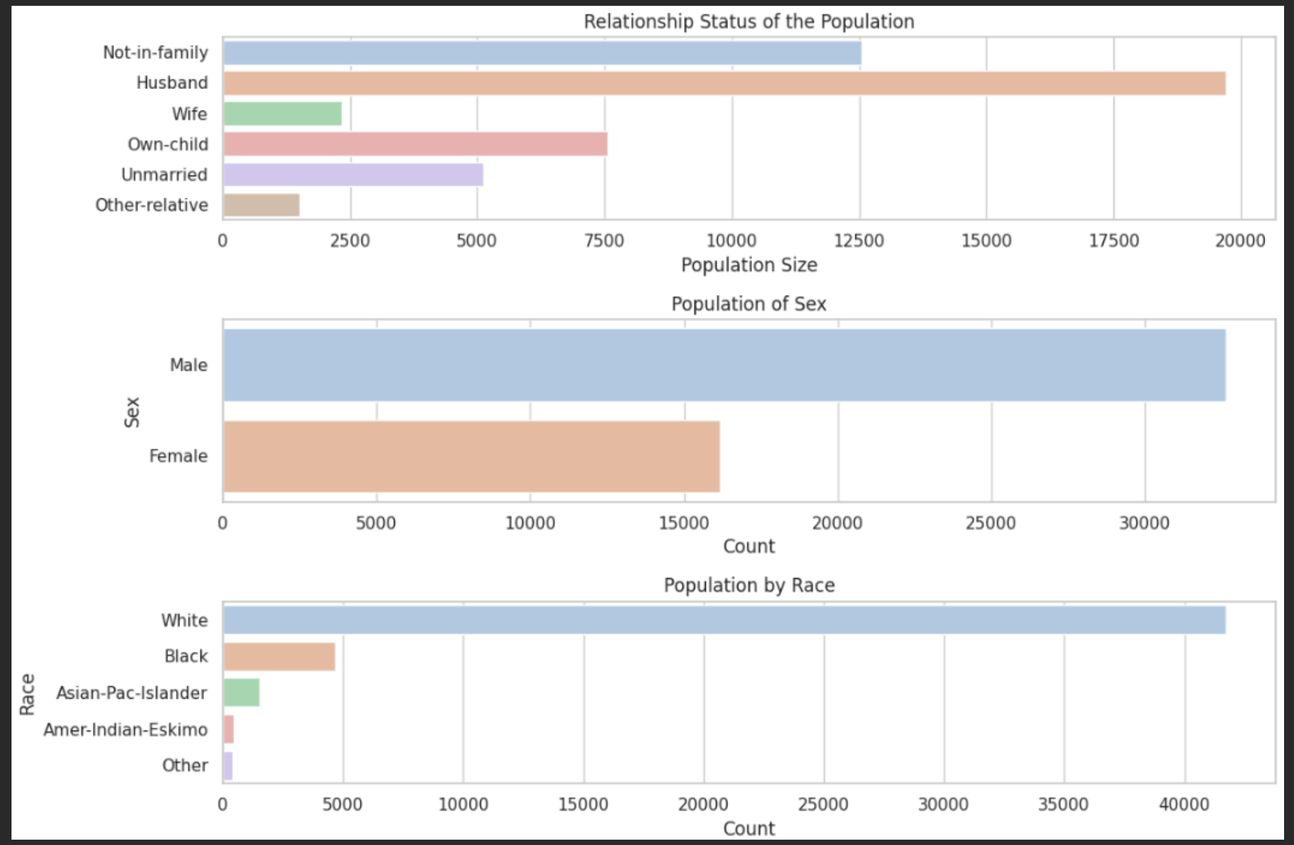
John louie V. Adornado



CENSUS INCOME

Kohavi,Ron. (1996). Census Income. UCI Machine Learning Repository. https://doi.org/10.24432/C5GP7S.

RELATIONSHIP, SEX, AND RACE :

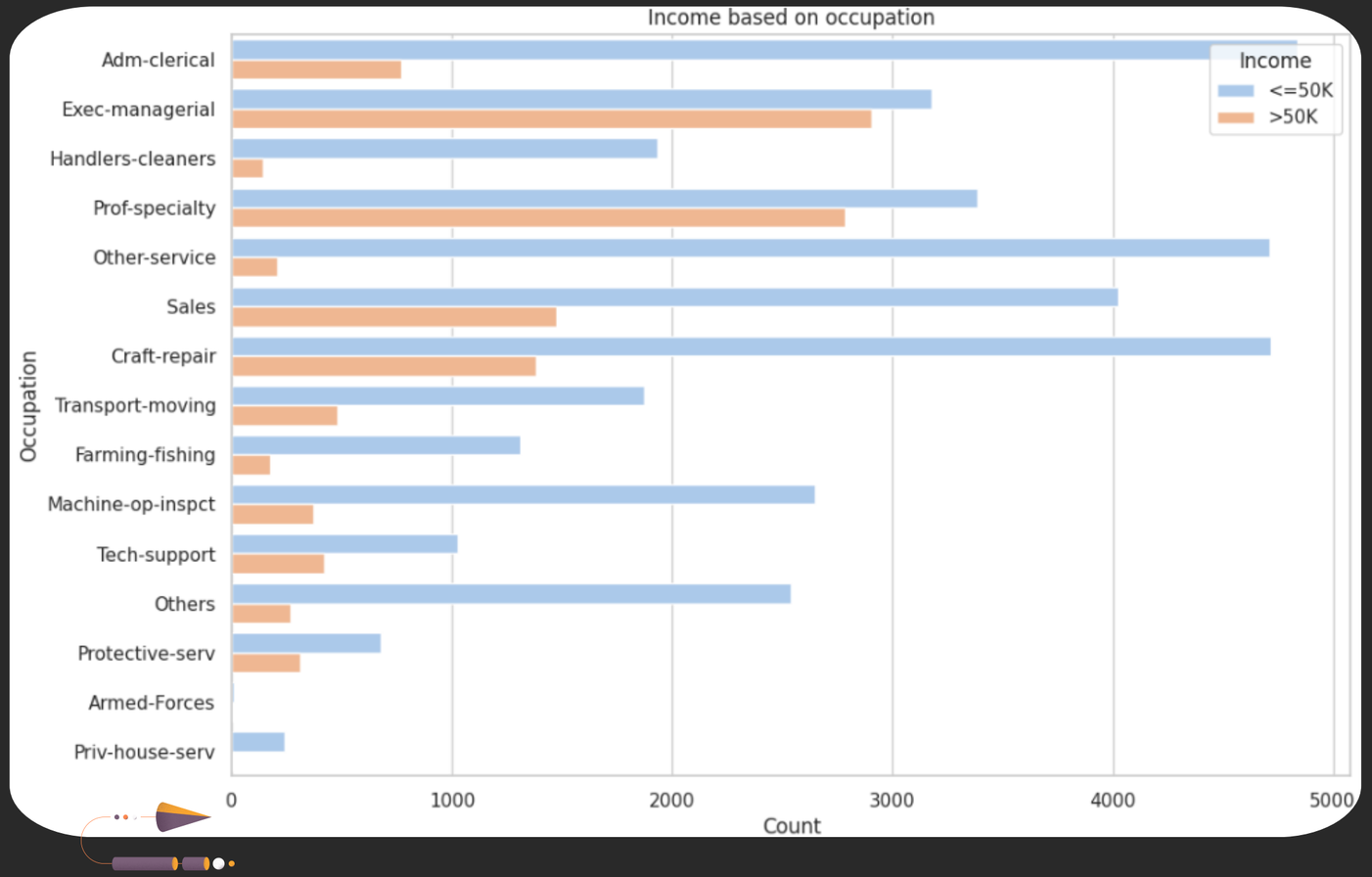


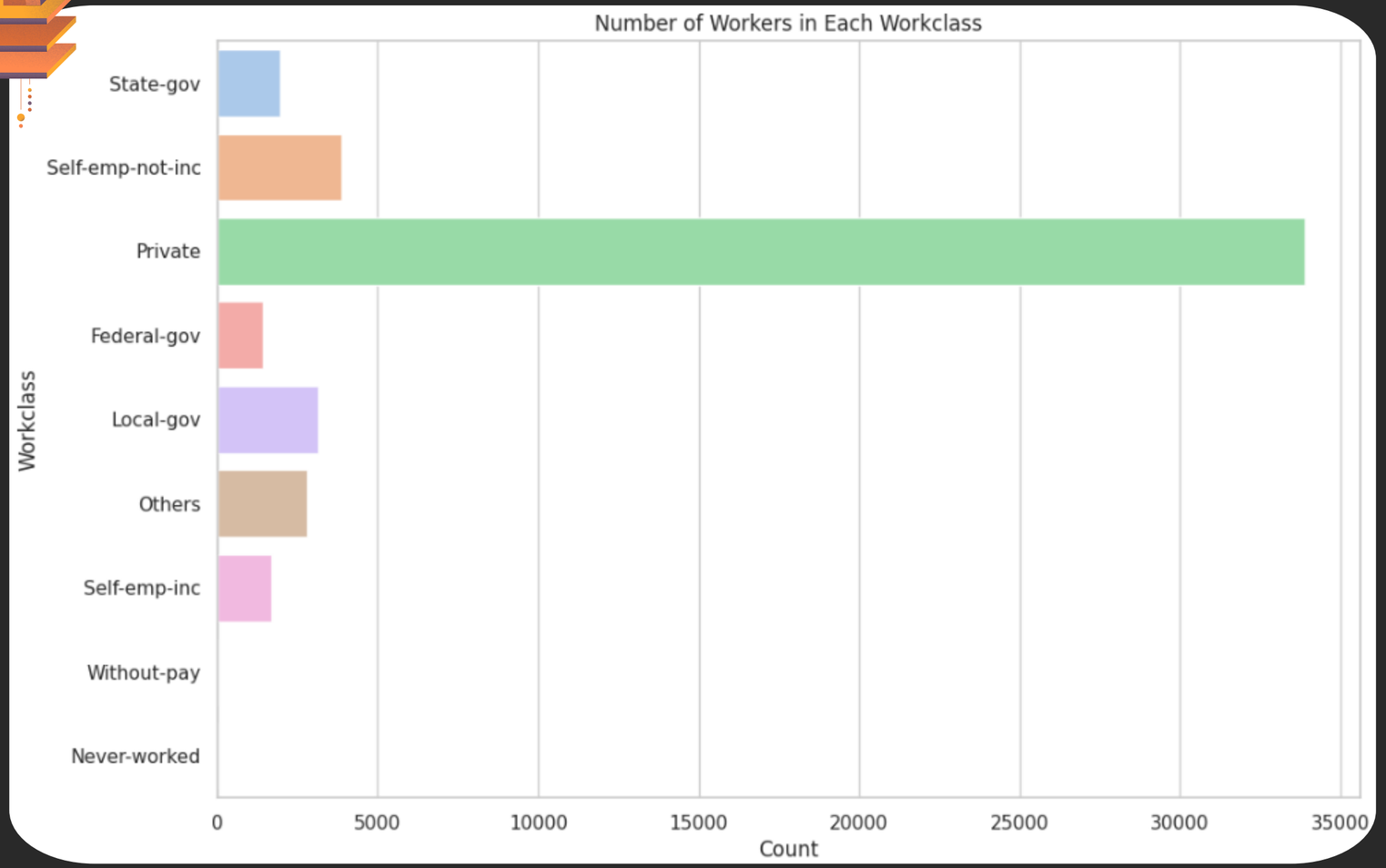
We're plotting the relationship status, gender, and race to understand the population distribution. This helps us see how many people are in relationships versus others. It also shows the distribution between males and females, as well as the racial composition of the population. These insights are valuable for understanding our dataset.

In summary, we will able to observe the correlation and difference of the population between people who are in a relationship, peoples sex, and peoples races.

INCOME BASED ON OCCUPATION

We're examining income across different occupations. It's evident that a majority earn less than or equal to 50k, outnumbering those earning more than 50k. For instance, in farming-fishing, a considerable number of individuals earn less than 50k, with only a small percentage earning above this threshold. This suggests a significant portion of individuals in each occupation earn lower incomes



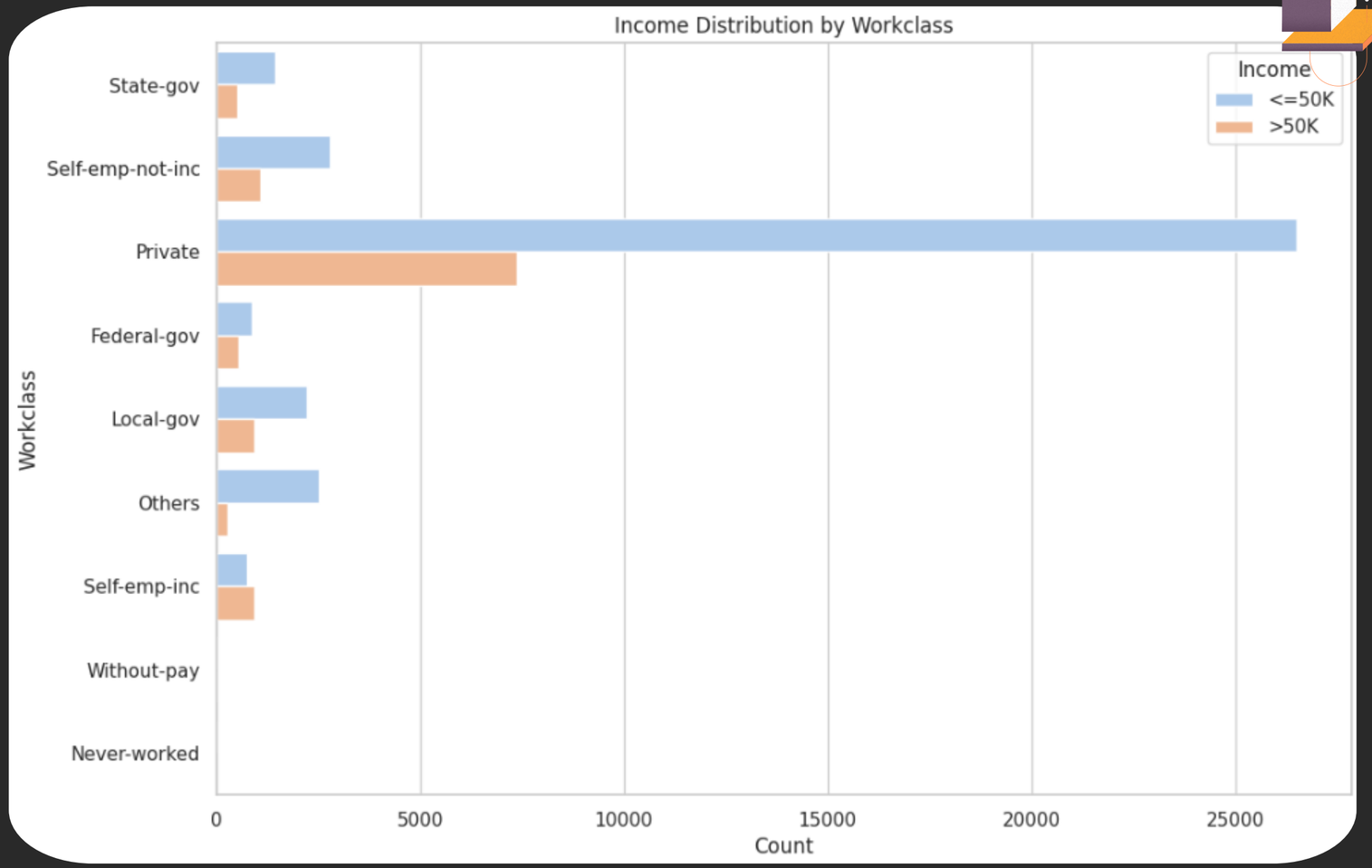


NUMBERS OF WORKERS IN EACH WORKCLASS

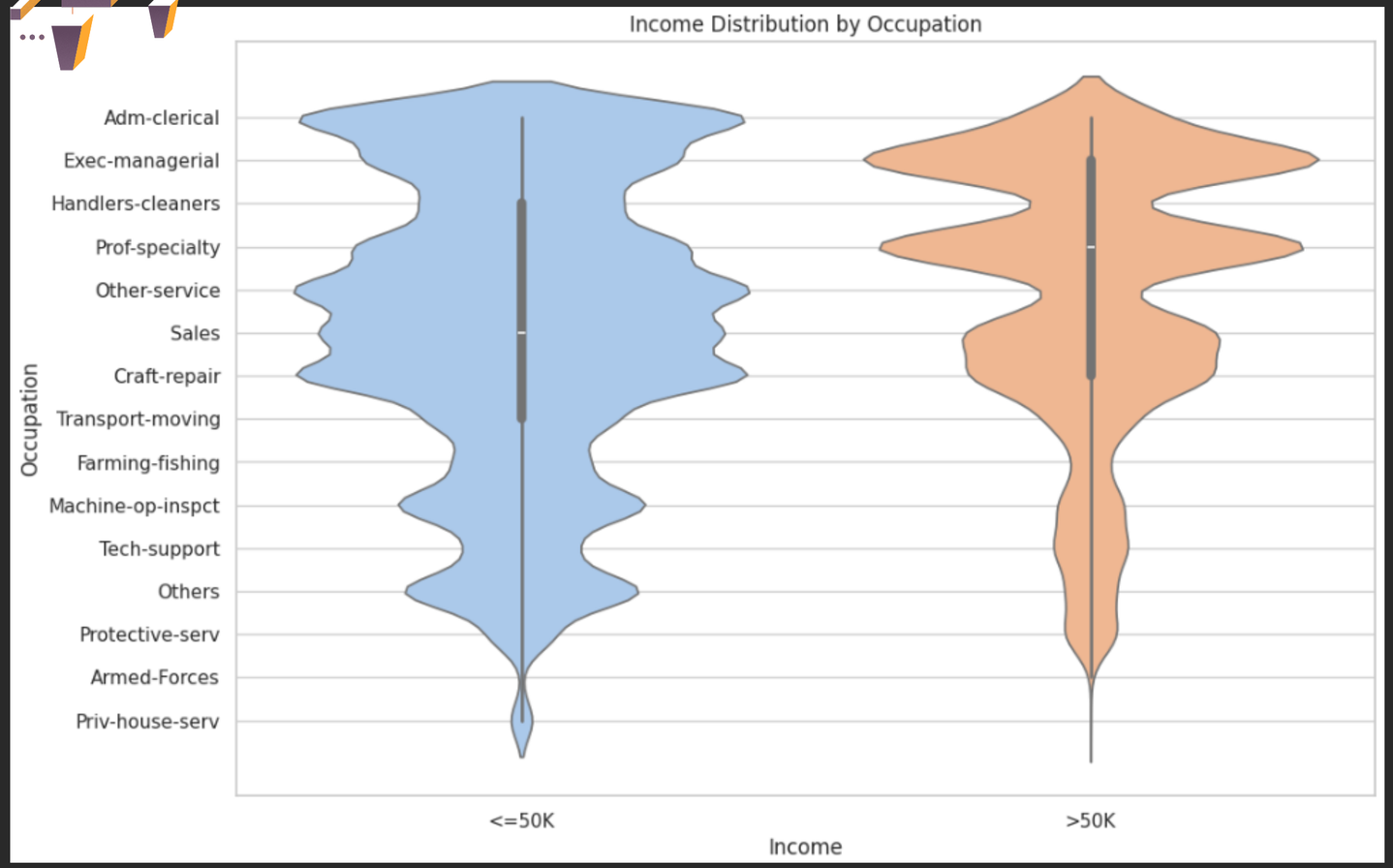
A large proportion of workers across various work classes are employed in the private sector. This indicates that a significant number of individuals, regardless of their specific work class, are engaged in employment within private companies or organizations.

INCOME DISTRIBUTION BY WORKCLASS

We're examining income distribution across different work classes. It's clear that individuals in the private sector dominate in numbers. Most of them earn less than 50k, but there's also a notable portion earning more than that. This higher-income group in the private sector still outnumbers those in other work classes.



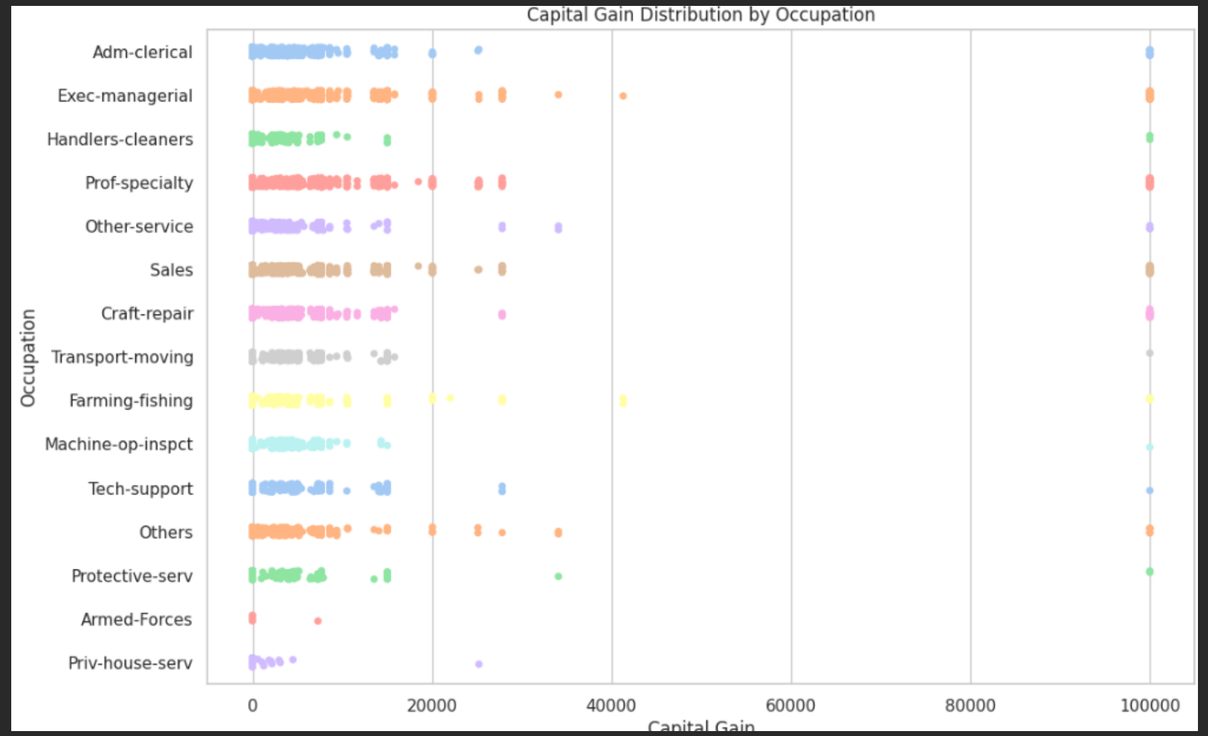
INCOME BY OCCUPATION



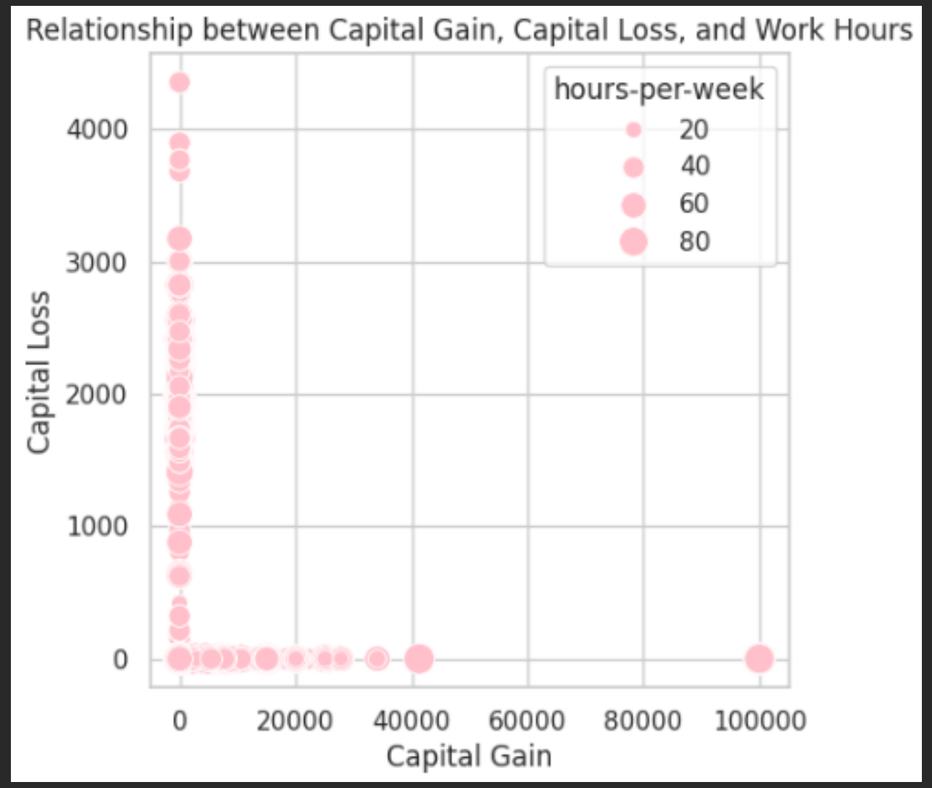
We've created a violin plot displaying the income distribution across different occupations. The violin plot provides a visual representation of the median income and other statistical measures, resulting in a more accurate depiction of the income density for each occupation. While it may be slightly more complex to interpret, it allows us to observe the density of income distribution within each occupation more effectively

CAPITAL GAIN

We've generated a strip plot to visualize how capital gains are distributed across different occupations. This plot offers a clear depiction of the range and distribution of capital gains within each occupation, providing insights how these gains vary across different professional fields.



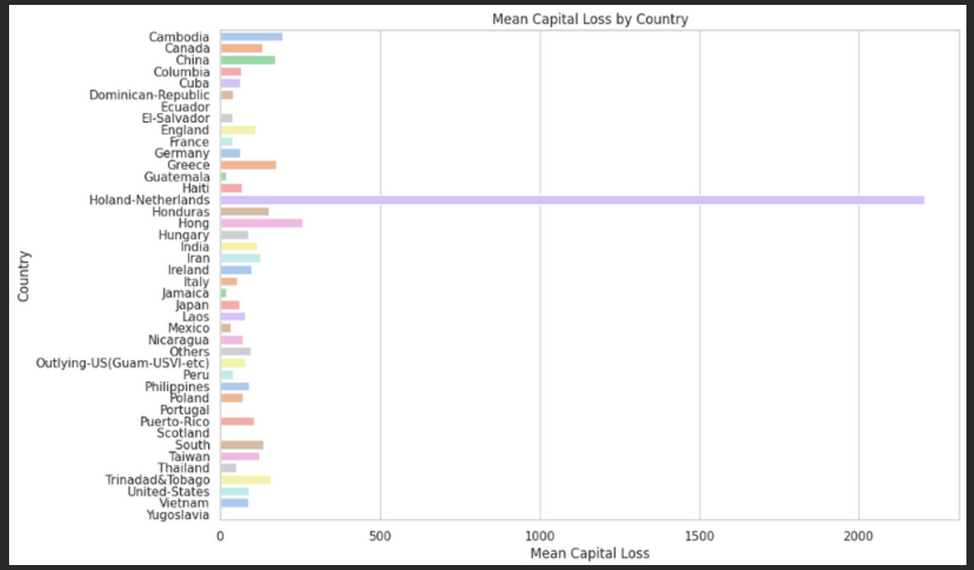
RELATIONSHIP BETWEEN CAPITAL GAIN, CAPITAL LOSS AND WORK HOURS

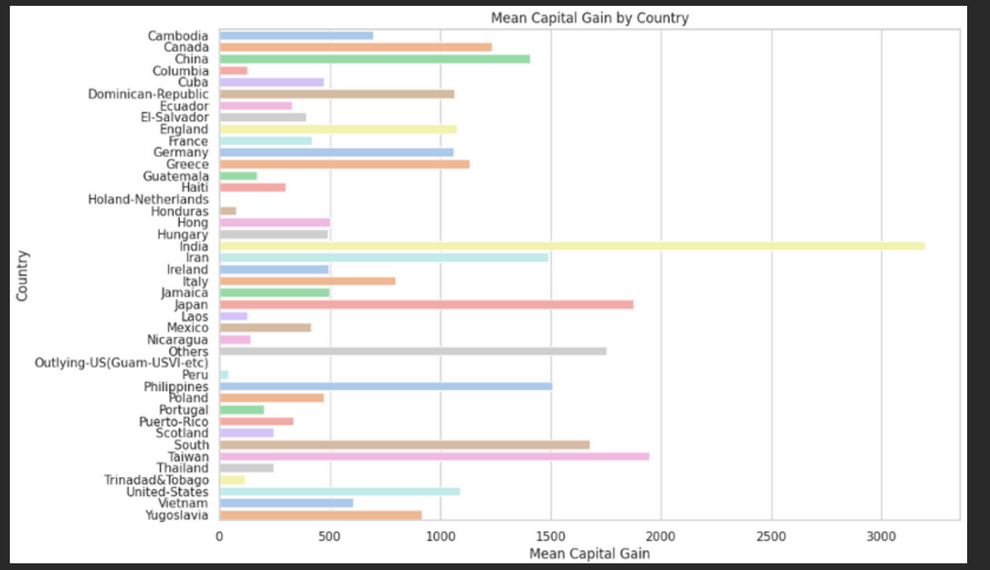


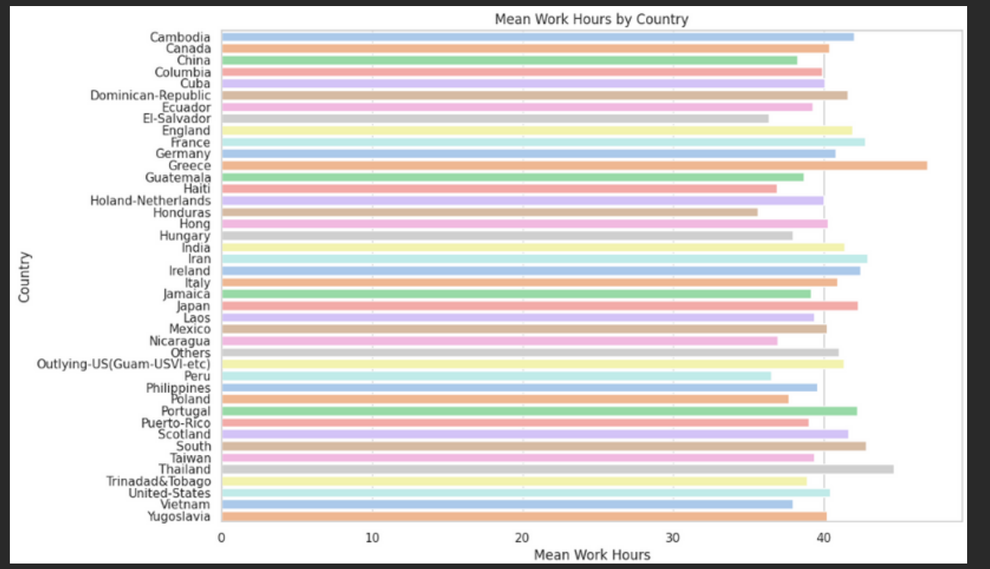
This also shows how the hours per week is shown by the size of it

We've presented a scatter plot illustrating the relationship between capital gain and capital loss, along with their respective counts, alongside the number of hours worked per week. The rationale behind exploring this relationship lies in understanding how work hours influence individuals capacity to generate the capital used for investment. Many individuals devote their time and effort to their careers or businesses to earn income, which forms the basis for their investments. The number of work hours directly impacts the amount of income one can generate, thereby influencing their ability to invest and subsequently realize capital gains or losses. By examining these variables together, we gain insights into the interplay between work hours, income generation, and outcomes

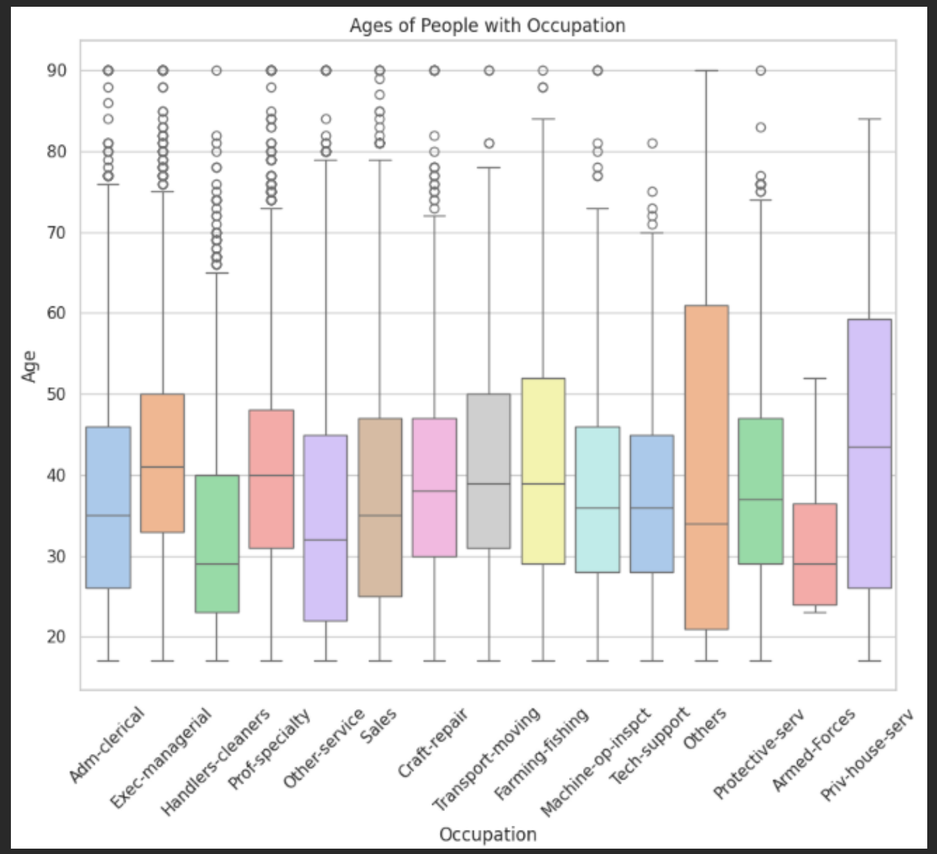
MEANS OF CAPITAL GAIN AND LOSS AND ALSO THE WORK HOURS OF EVERY COUNTRY







Upon analyzing the plot, it becomes evident that India has the highest mean capital gain among the observed countries. Additionally, the mean capital loss for each country highlights that the Netherlands surpasses all others in this aspect. Furthermore, in terms of work hours, the mean across nearly all countries hovers around 40.



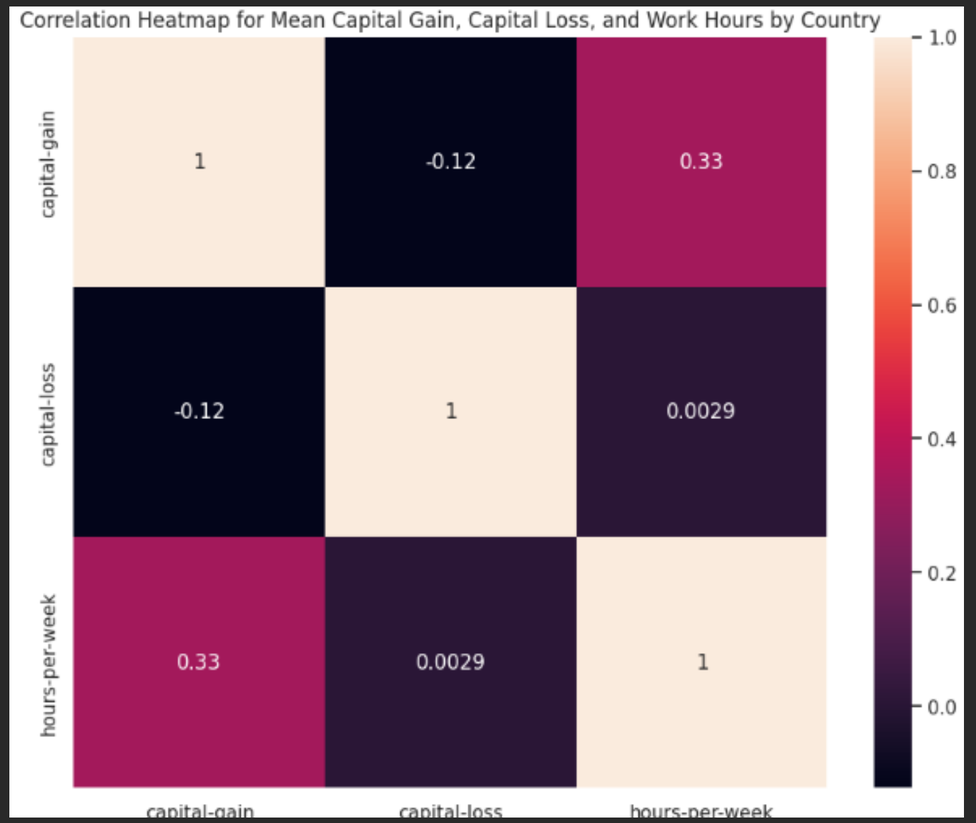
AGES OF PEOPLE WITH OCCUPATION

With these box plot of ages of people with occupation we see that there is many outliers (Ages of people who surpasses the populations age) and the “others” have the max value surpasses others and it also surpasses everyone in the 75th percentile and the mean percentile of the armed forces is not the same as others while others are in the below 20 the armed forces is at above 20 below 30

CORRELATION HEATMAP

- The correlation coefficient between Hours Per Week and Capital Gain is 0.33. This indicates that there is a weak positive correlation between the number of hours worked per week and the amount of capital gain. In other words, as the number of hours worked per week increases, there tends to be a slight increase in capital gain, although the relationship is not very strong.

- The correlation coefficient between Hours Per Week and Capital Loss is 0.0029. This suggests that there is a very weak positive correlation between the number of hours worked per week and the amount of capital loss. The correlation is almost negligible, indicating that there is little to no relationship between these two variables.



- THE CORRELATION COEFFICIENT BETWEEN CAPITAL GAIN AND CAPITAL LOSS IS -0.12. THIS INDICATES A VERY WEAK NEGATIVE CORRELATION BETWEEN THE AMOUNT OF CAPITAL GAIN AND THE AMOUNT OF CAPITAL LOSS. IN OTHER WORDS, AS CAPITAL GAIN INCREASES, THERE IS A SLIGHT DECREASE IN CAPITAL LOSS, AND VICE VERSA. HOWEVER, THE CORRELATION IS QUITE WEAK.

DATA ANALASYS SUMMARY & CONCLUSION

John louie V. Adornado

In my analysis of the "Census Income" dataset, I examined incomes based on occupation, relationship status, and compared incomes across different workplaces. I observed significant differences in workplace distributions, with the private sector showing a higher density in the graph, as indicated by the dataset. Income levels are strongly influenced by both the type of work and the position held. For instance, even individuals engaged in farming or fishing businesses can earn substantial incomes, surpassing those in other occupations. Furthermore, I identified which occupations or positions yield the highest incomes.

In conclusion, the analysis of various visualizations and statistical measures provides valuable insights economics of different occupations. Across the dataset, the distribution of income, capital gains, and work hours varies significantly among occupations, with notable differences observed across countries as well. While some occupations exhibit stronger correlations between certain variables, others show weaker or negligible relationships. Moreover, the presence of outliers in age distributions highlights the diversity within each occupation. Overall, these findings underscore the complexity economic dynamics, which shows the importance of considering multiple factors when analyzing work dynamics.