CSE578 – Data Visualization – Final Report

Krithiga Venkataraman, *krithiap@gmail.com*

Abstract

Analyze United States Census Bureau data to identify factors which determine an individual's income. These factors will help in creating marketing profiles to help bolster college enrollment. Demographic data is a powerful tool which when properly cleansed and analyzed can provide rich insights. The analysis process involves cleansing data, identifying key attributes which will be considered for analysis, designing scenarios and then creating plots to help visualize the data. The plots are then analyzed to derive key insights to draw conclusions. I have used univariate and multivariate plots for this analysis. Histogram univariate plot is used to analyze distribution of age. Multivariate plots used include mosaic plots to analyze gender and income variables and as well as work class and income, a parallel co-ordinate plot to analyze gender, age, education and income. I have also used scatter plot to analyze education and capitalgain and finally a box plot to analyze race and capitalgain. These plots have helped tremendously to identify trends and relationships which are otherwise difficult to spot just looking at the data. The plots have helped derive conclusions to help answer the core question, which is to identify variables which help determine an individual's income.

I. Introduction

As a data analyst working for XYZ, I need to create marketing profiles using salary as a key demographic leveraging data supplied by the United States Census Bureau so that I can help UVW college to bolster its enrollment.

The main goals are to

1. Find factors which determine an individual's income.
2. Design and implement Univariate and Multivariate Data Visualization techniques to analyze & recommend which variables impact the income the most.

A. Overall approach

1. Analyze and understand the dataset.

Understand the structure of the United States Census data by describing its schema. Analyzing the schema to understand which columns categorical and which columns are continuous, will help in choosing the right chart and model for prediction. Querying the data to understand the unique values for different categorical columns. Analyze input data using Pandas python library and leveraging functions like group by, filter and summarizations to analyze key factors.

2. Identify the variables to be used in the visualizations.

Using the features analyzed in task 1, I identified eight key variables which will be used in univariate and multivariate charts.

These variables are:

a. age,
b. workclass
c. education (& education-num)
d. marital status
e. race
f. gender

g. capital-gain
h. hours-per-week.

3. Design scenarios & plot visualization charts to help answer the core problem statement.

Design the scenarios which will analyze the relationships between the variables identified in task B. Univariate or multivariate charts will be created for each scenario to bring out the relationships visually.

4. Derive conclusions.

Finally, I need to derive conclusions and describe how the each of the scenario and the chosen factors affects the income level. The identified relationships will help the marketing team to come up with targeted marketing material which will help UVW college to bolster its enrollment.

B. Assumptions

These are the assumptions made for this analysis.

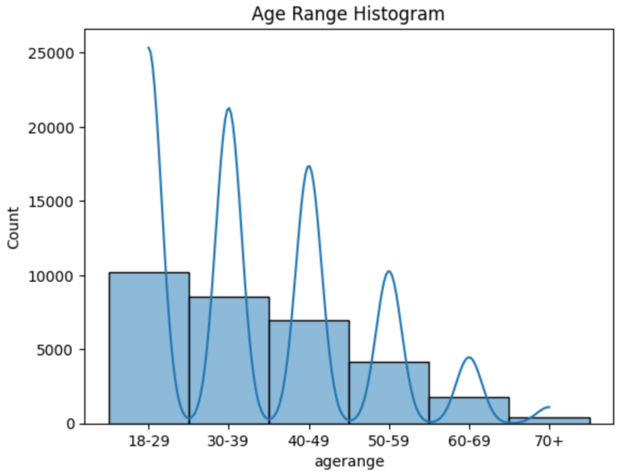
- Data used from public source is accurate and reliable.
- Capitalgain is a good proxy for income.

C. Description of scenarios

This project required analyzing the relationships between key variables from the United States Census Bureau data to help establish patterns which will decide the marketing strategy. The project will analyze the relations through these five scenarios with supporting Data Visualizations for each.

1. *The marketing team wants to know the distribution of age based on age ranges.*
   The raw data has age as a continuous variable, I decided to create a new variable called age range to create a categorical variable of age ranges in 10-year intervals. This will help analyze how many people are present in each age interval. I then proceeded to create a histogram plot to analyze the distribution of people based the age range to decide which age group should be targeted for enrollment.

   The histogram plot clearly shows that the distribution of people in different age ranges and that we should focus on the 18-49 age range which has the highest distribution of people and the age group which would be most interested in a degree to earn a higher wage.

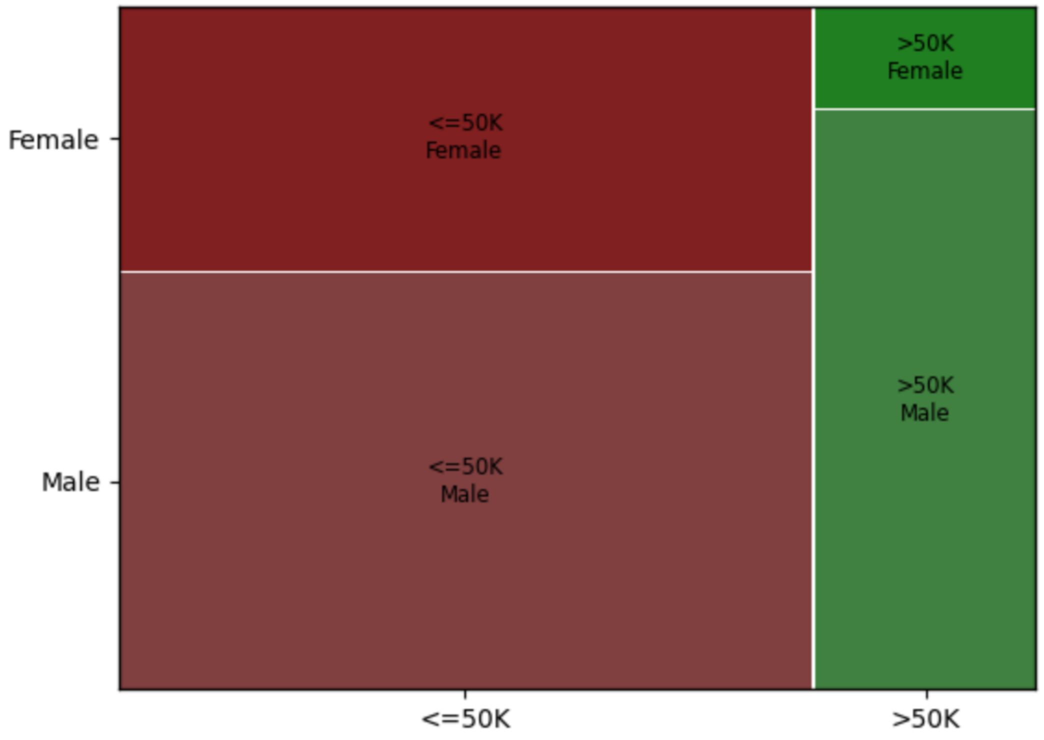


2. Analyze how gender and income are related to each other.

To analyze how gender and income are related, I decided to use a mosaic plot to understand their relationship.

The mosaic plot shows that about 70% of the population earn less than 50K and only about 30% earn more than 50K. The chart also shows that the percentage of male's earning above 50K is significantly higher than the female population. The distribution of male to female earning less than 50K is approximately 60% to 40%.

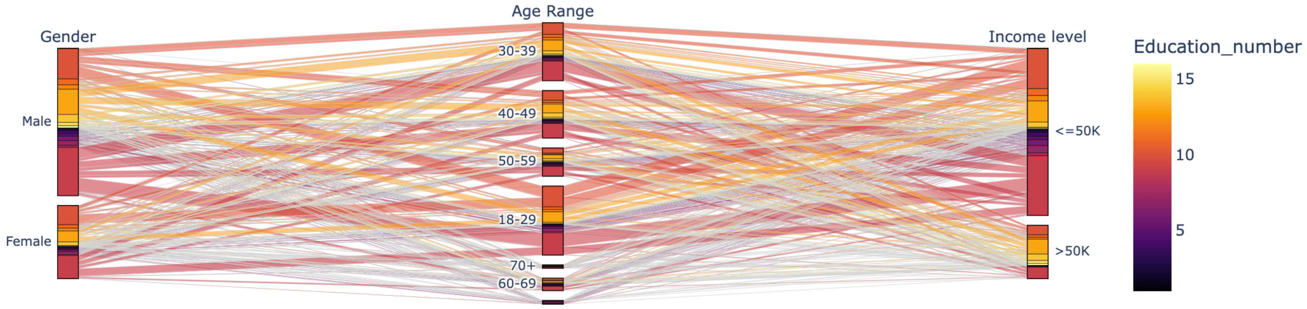
As we can visualize, there is clearly a relationship between income and earning ability as male population seems to generally earns more than females the disparity is even more evident in the group earning more than 50K. This seems to suggest than we can use this variable to target female & male population earning less than 50K.



3. Analyze how level of education, capital gain and age impacts income levels.

The marketing team is also interested to analyze the relationships between multiple variables like level of education, capital gains, age and their impact on income level. I decided to use a parallel co-ordinate plot to analyze this relationship.

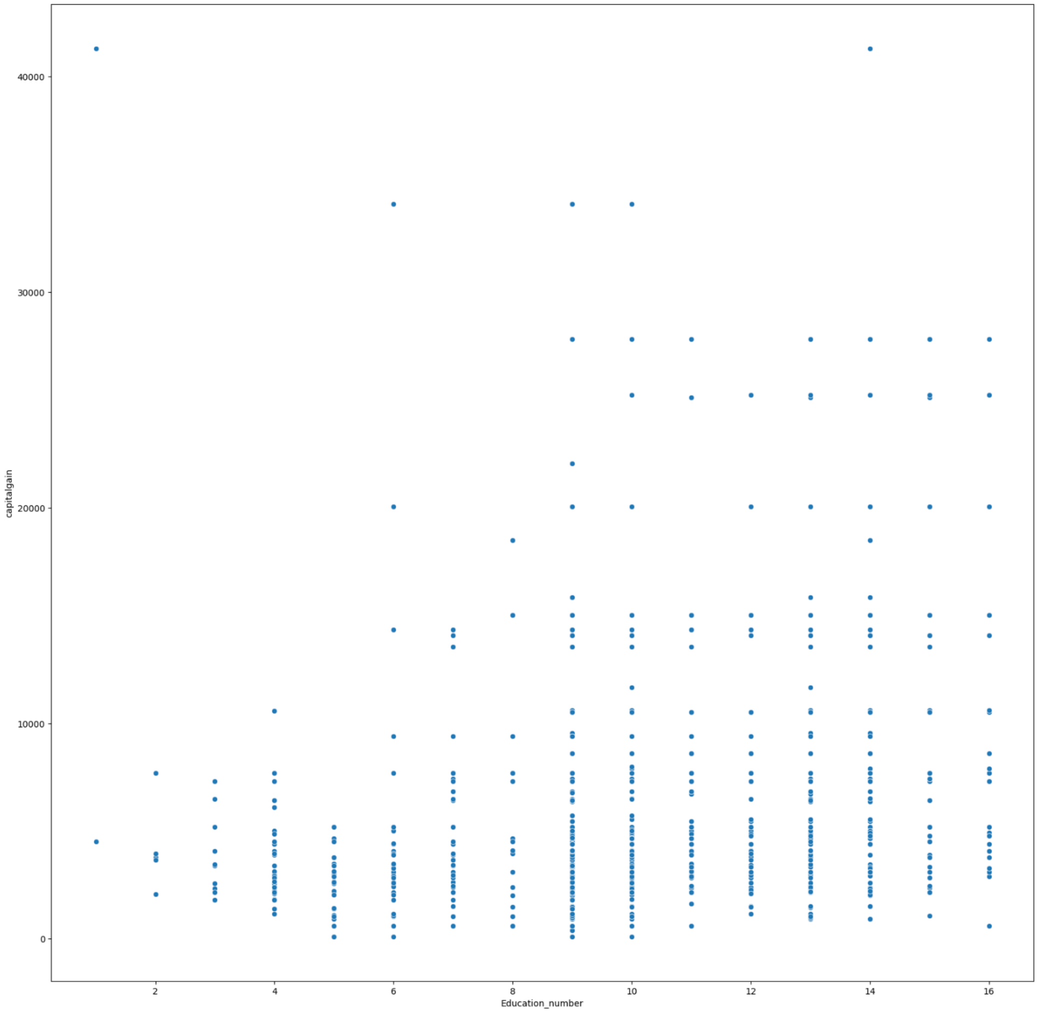
A parallel co-ordinate plot will come in handy to explore this relationship and to bring out how these variables are connected to each other.



4. Analyze the relationship between level of education and capital gain.

Another interesting relationship to explore would be how level of education affects capital gain. This would be a key data point to understand the correlation between level of education and how it directly impacts the financial stability as this would help design a strong marketing strategy.

To explore this relationship, I will be using a scatter plot to visually represent the relationship between these two variables.

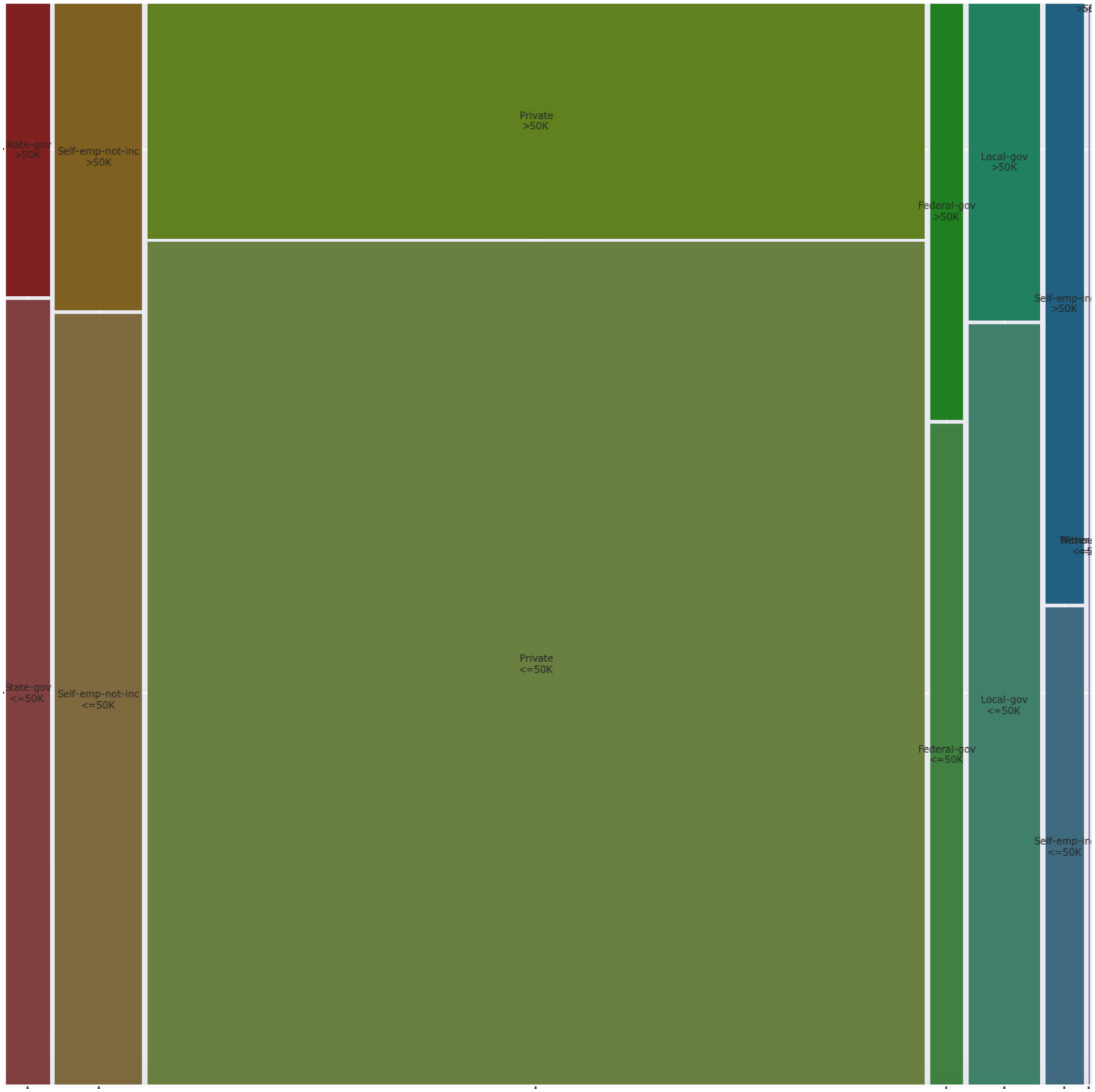


5. How are workclass and income related?

One other question the marketing team posted is how workclass affects income. This is again an important relationship to understand so that the team can focus its effort to design appropriate marketing material for the target population.

To analyze this relationship, I chose a Mosaic plot to visualize this relationship.

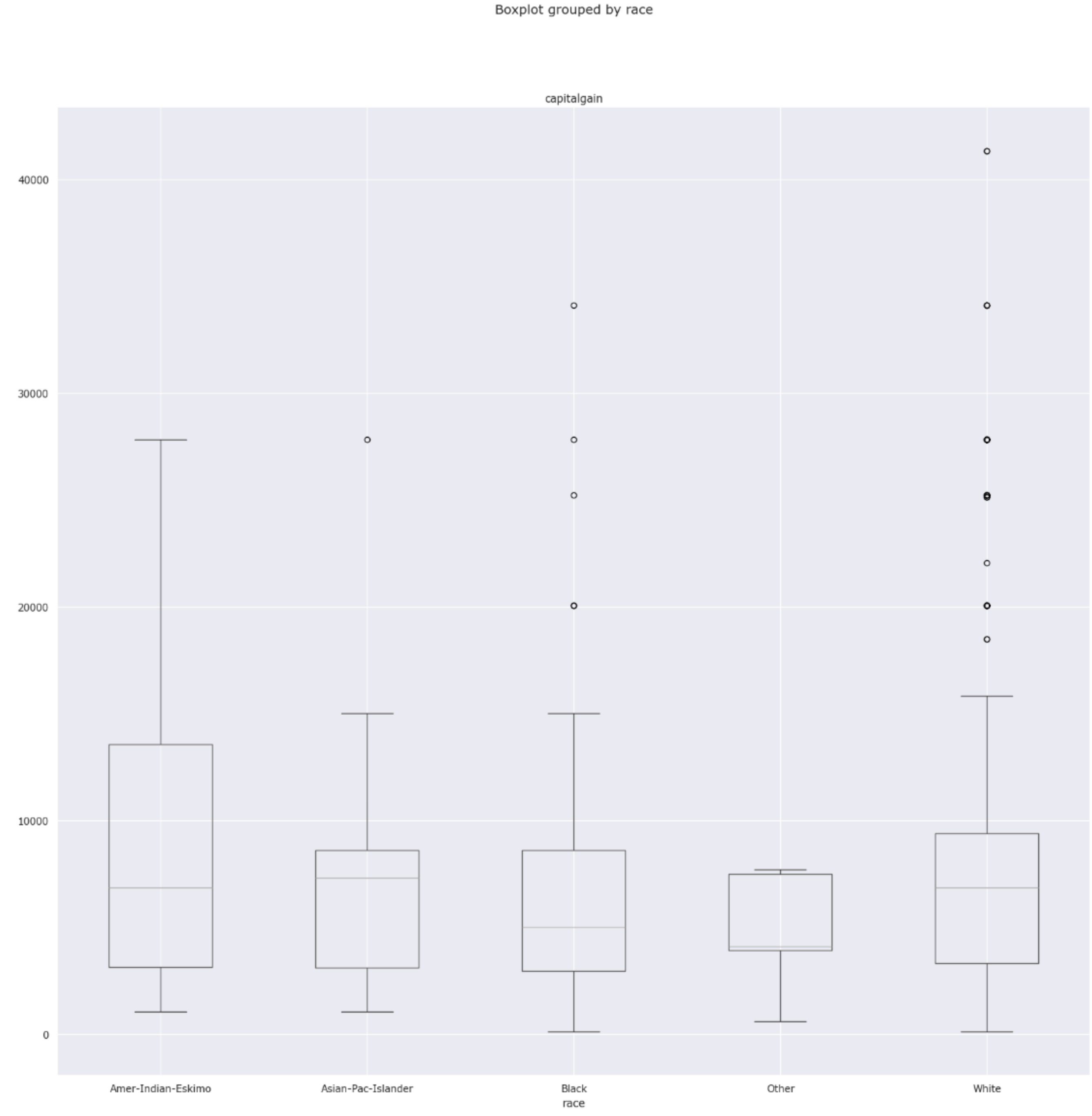
The plot shows that a substantial section of the population employed in private sector earn under 50K



6. How does race affect capitalgain?

One other question the marketing team posted is how race affects capitalgain, which can be considered as a proxy variable for income. This is again a key relationship to understand to come up with appropriate marketing material.

To analyze this relationship, I chose a Boxplot to visualize this relationship. One interesting observation from this plot is that the American-Indian-Eskimo group seem to have a much wider range of capitalgain with skew towards the higher end of the spectrum.



II. Conclusion

In conclusion, the variable I chose for analysis have different impacts on income, with age, gender, level of education and race having most impact on income. The histogram plot of the data based on age shows that significant age of the population is under 49 and this group is a good candidate for marketing education. The mosaic plot clearly brings out the income disparity between male and female genders with males generally having a higher income than females. The scatterplot between level of education and capitalgain confirms the general assumption that income increases with level of education and so is a good metric to judge the income of a person. The boxplot brings out interesting facts on the race and income level, which again can be used to design specific campaigns to help certain sections of the population to earn a better income and living.