Group 8

Project : PREDICTION OF FACTORS INFLUENCING SALARY

Agenda

- Introduction
- Visualizations
- Conclusion
- Project Contribution

Introduction

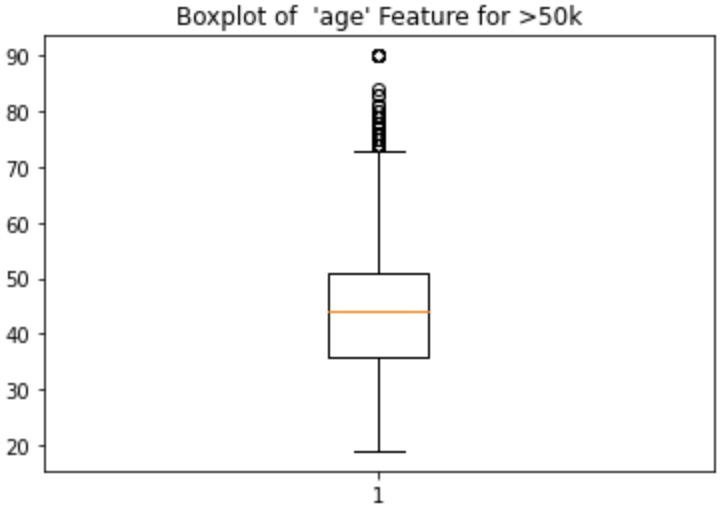
Our project aims to develop marketing profiles for an XYZ corporation where we work as Data Analysts. We develop different marketing profiles for UVW college where we derive our input data from the United States Census Bureau. Our key to developing marketing profiles is based on the Salary/income of each individual based on different values of the input parameters, which would help the corporation to directly target those based on their income. We also group the factors that can be used in the development of their proposed model/application.

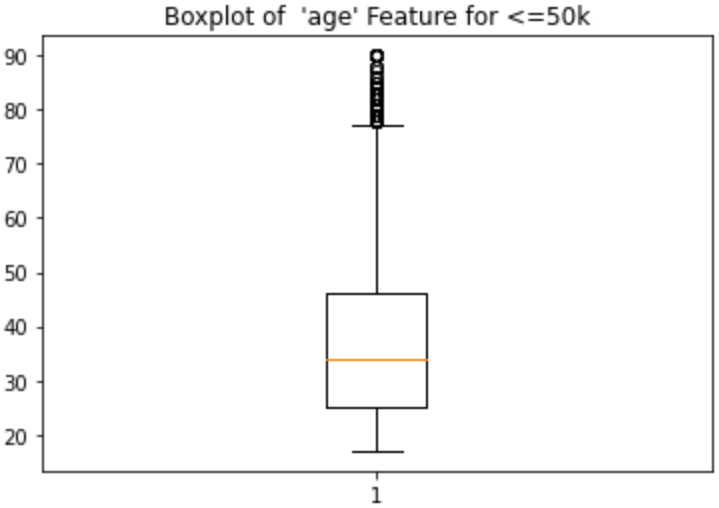
Visualizations

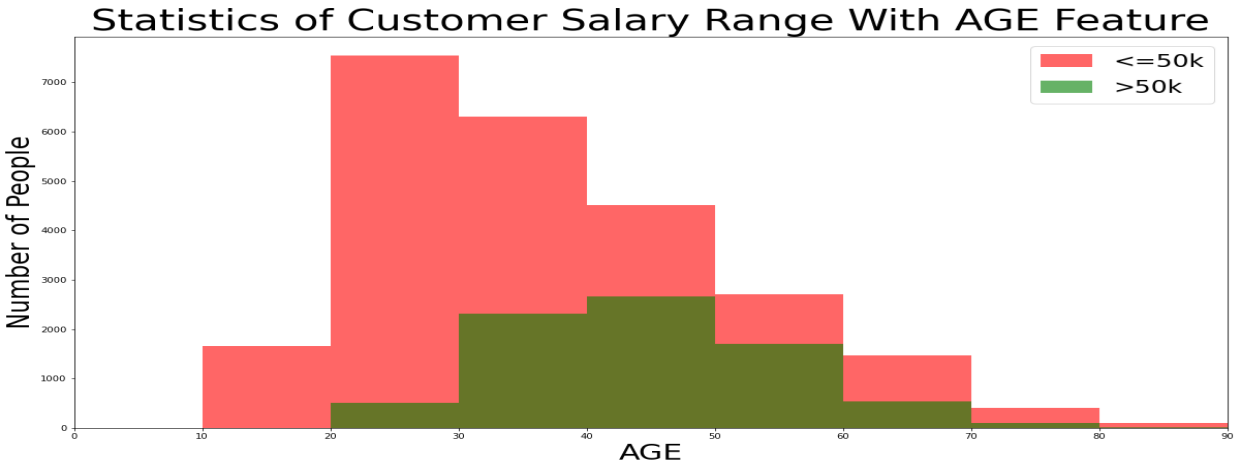
Age

As age increases, the number of individuals with a salary >50k increase up to a certain age from the above overlapping histogram, whereas the number of individuals with a salary <=50k decreases up to a certain age.

<u>Inference</u> -
Hence, XYZ corporation can use 'age' as a feature for their prediction model.



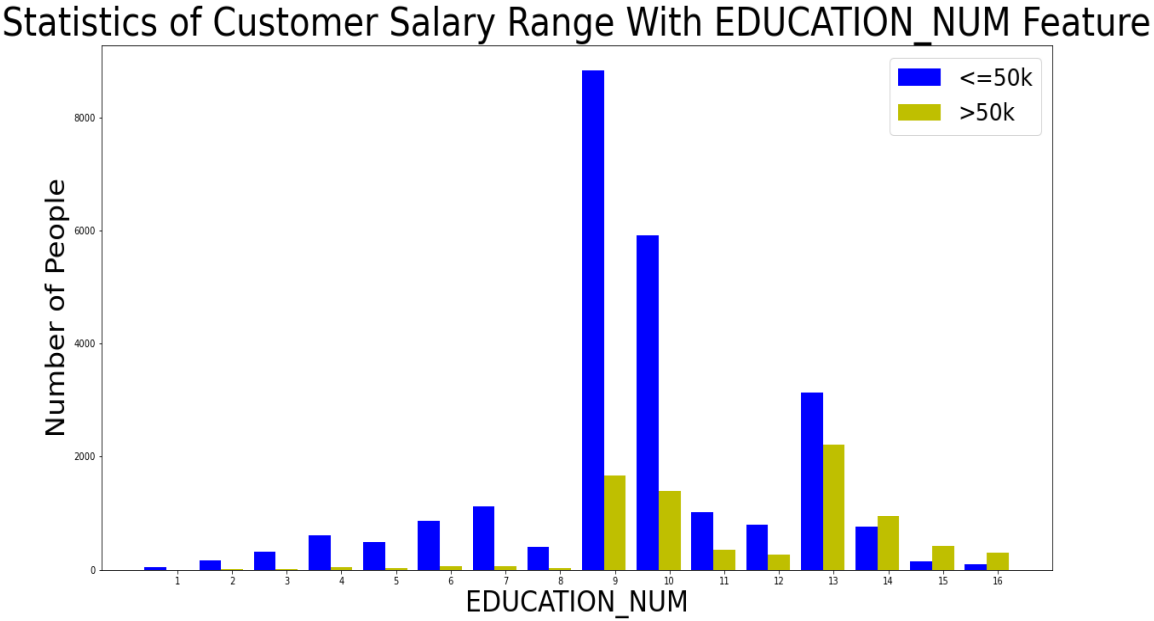


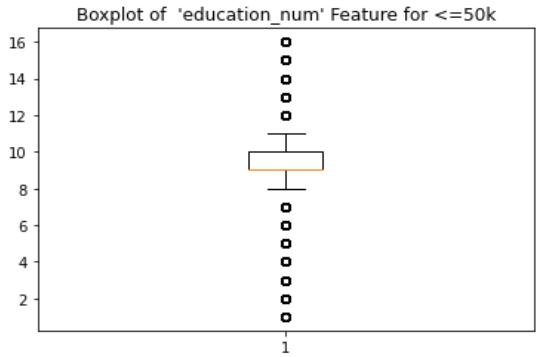


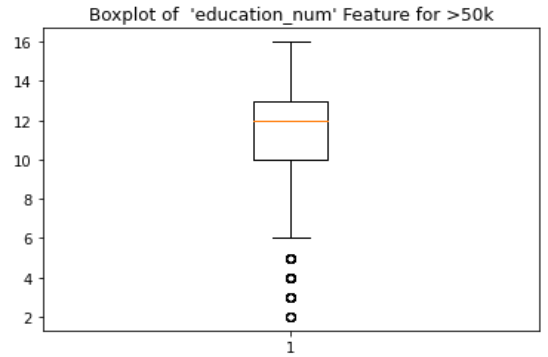
Education Number

The visualization suggests that people with education num 9 and 10 would highly likely earn a salary <=50K.

<u>Inference</u> - The visualization suggests that people with education num 9 and 10 would highly likely earn a salary <=50K.Hence, XYZ corporation can use Education Num as a feature for their prediction model.



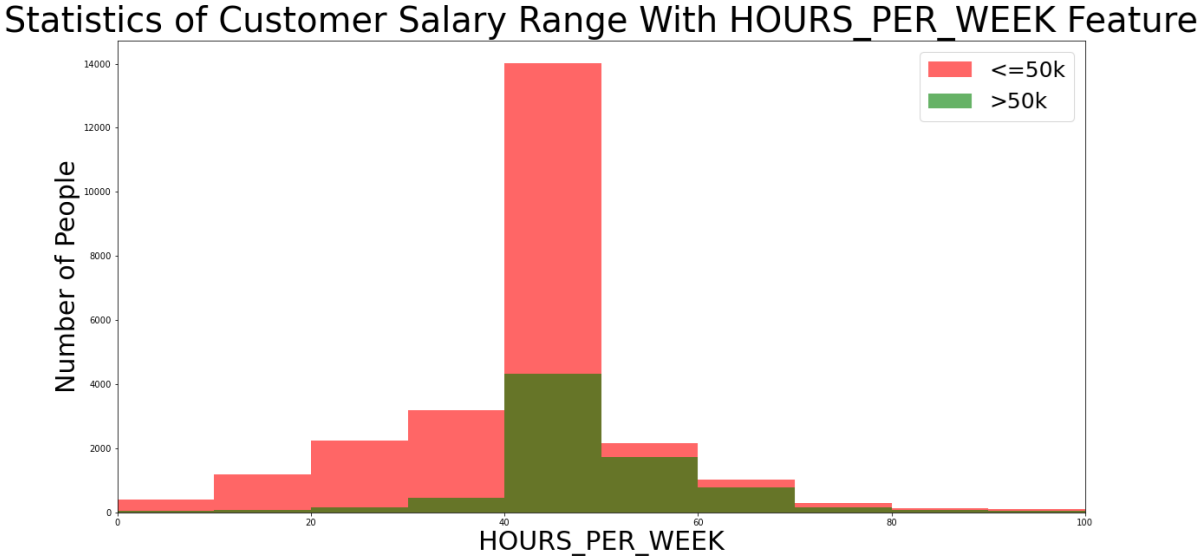


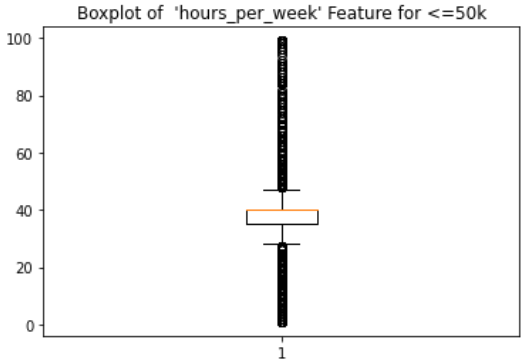


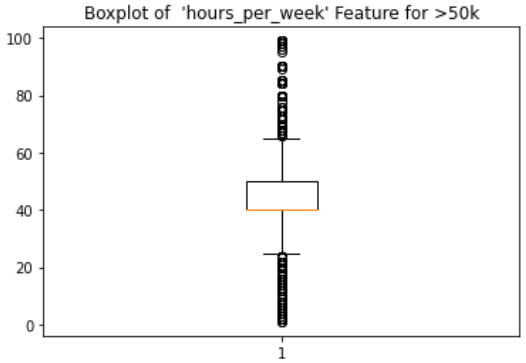
Hours Per Week

Since Hours Per Week is a continuous numerical variable, a histogram is a great way to visualize the pattern. We ought to equate data with salaries of >50k and <=50k; therefore, a stacked bar chart is more convenient. Overlapping histograms enable viewers to compare data quickly, and they also fit well for a wide variety of data.

<u>Inference</u> - The visualization suggests that the number of individuals who work for 40-50 work-hours/week has high chances of earning >50k and compared to others who make>50k.
Hence, XYZ corporation can use Hours Per Week as a feature for their prediction model.





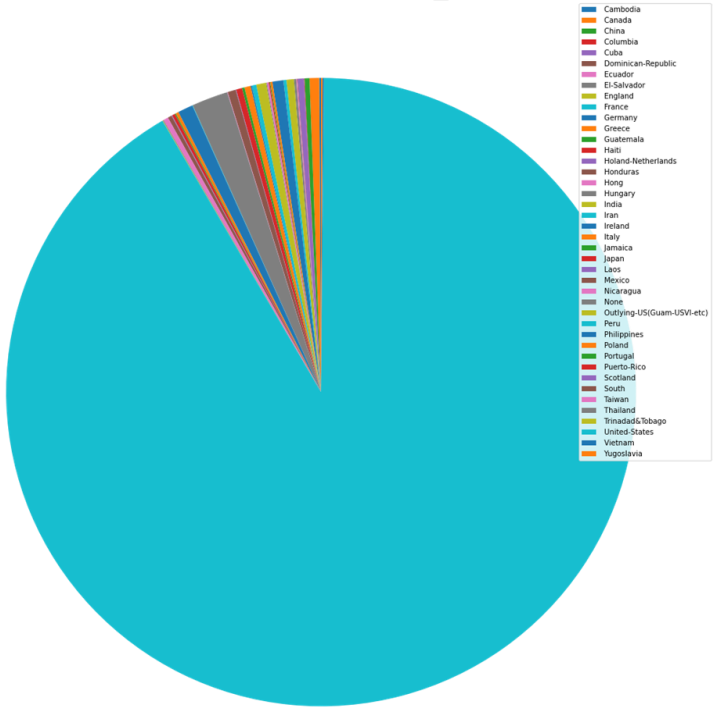


Native Country

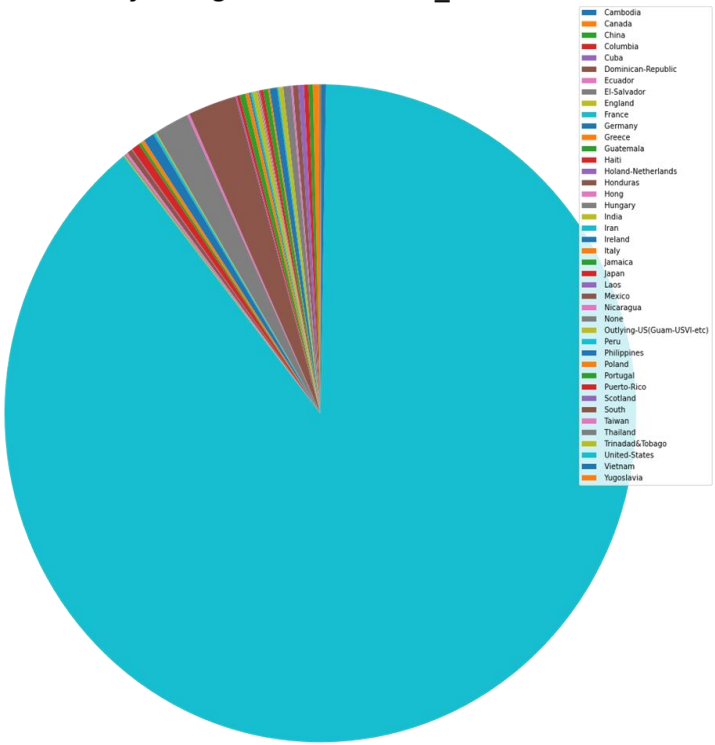
The Native Country feature is discrete categorical data with many categories. Pie charts are good at relative proportions of multiple data classes and summarize an extensive data set in the visual form; thus, we have used Pie charts for Native Country.

<u>Inference</u> - There are many individuals from the United States whose salary > 50k.

Statistics of Customer Salary Range With NATIVE_COUNTRY Feature for above 50k



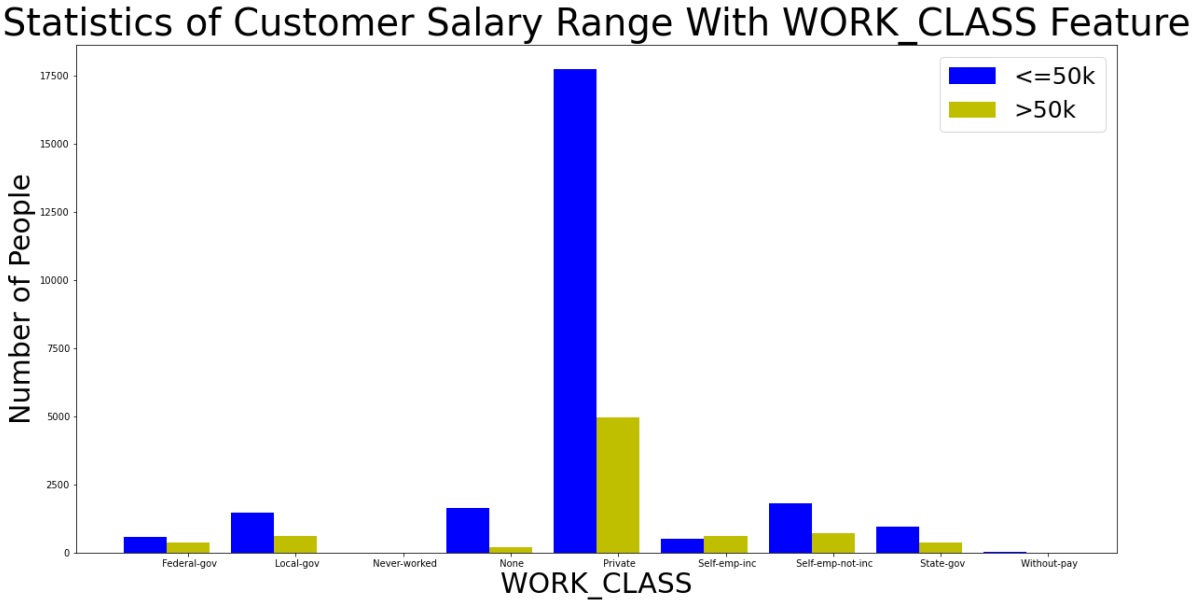
Statistics of Customer Salary Range With NATIVE_COUNTRY Feature for below 50k



Work Class

Work Class is a discrete categorical data. A bar graph with two different bins for salaries >50k and <=50k visualizes central values briefly.

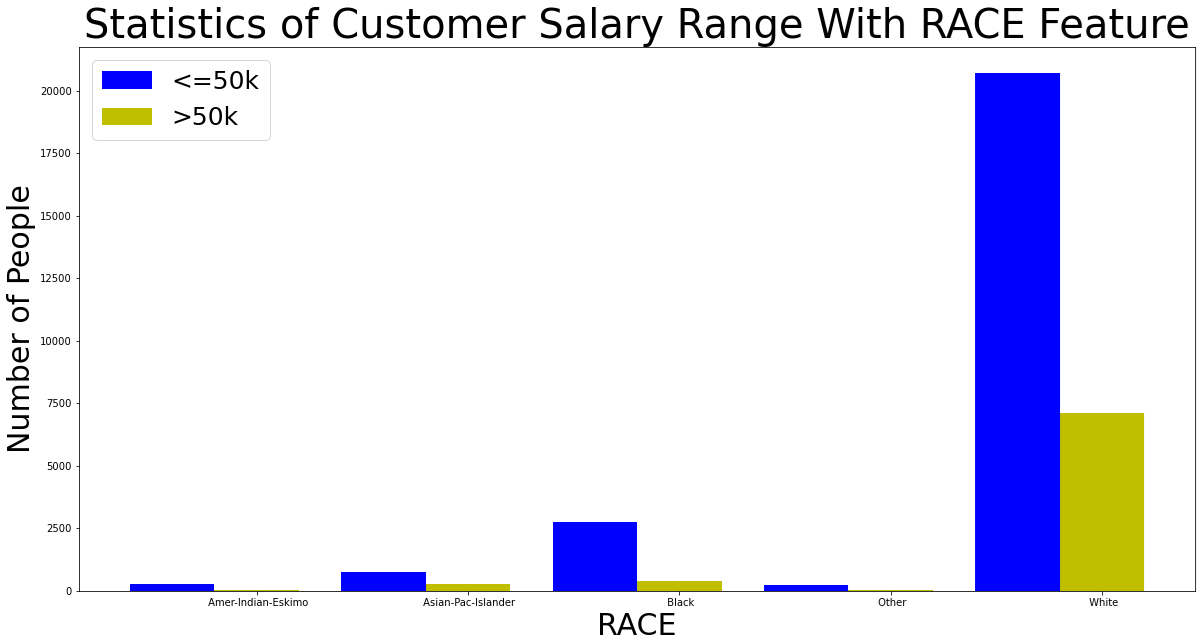
<u>Inference</u> - we can infer that individuals with a work class of "Private" are more likely to earn a salary <= 50k.
Hence, XYZ corporation can use Work Class as a feature for their prediction model.



Race

The race is discrete categorical data. A bar graph with two different bars for salaries >50k and <=50k visualizes central values at a glance, allowing a visual check of the correctness and reasonableness of estimate. By looking at the above histogram, the data is left skewed at the race “White”

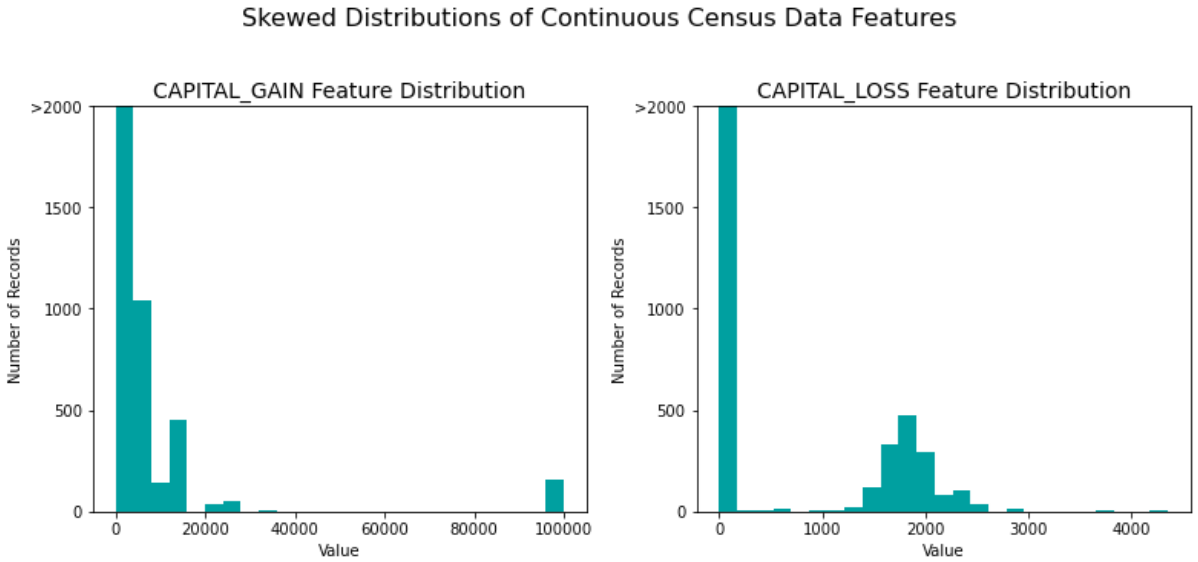
<u>Inference</u> - We can infer that individuals with feature class of Race as “White” are more likely to earn a salary > 50k
Hence, XYZ corporation can use race as a feature for their prediction model.



Capital Loss, Capital Gain

We have used Histograms for Capital loss, and Capital Gain features are continuous numerical variables.

<u>Inference</u> - There is a strong indication that people with capital gains between 1k - 20k are likely to earn >50k. Hence, XYZ corporation can use Capital Loss, Capital Gain as a feature for their prediction model.



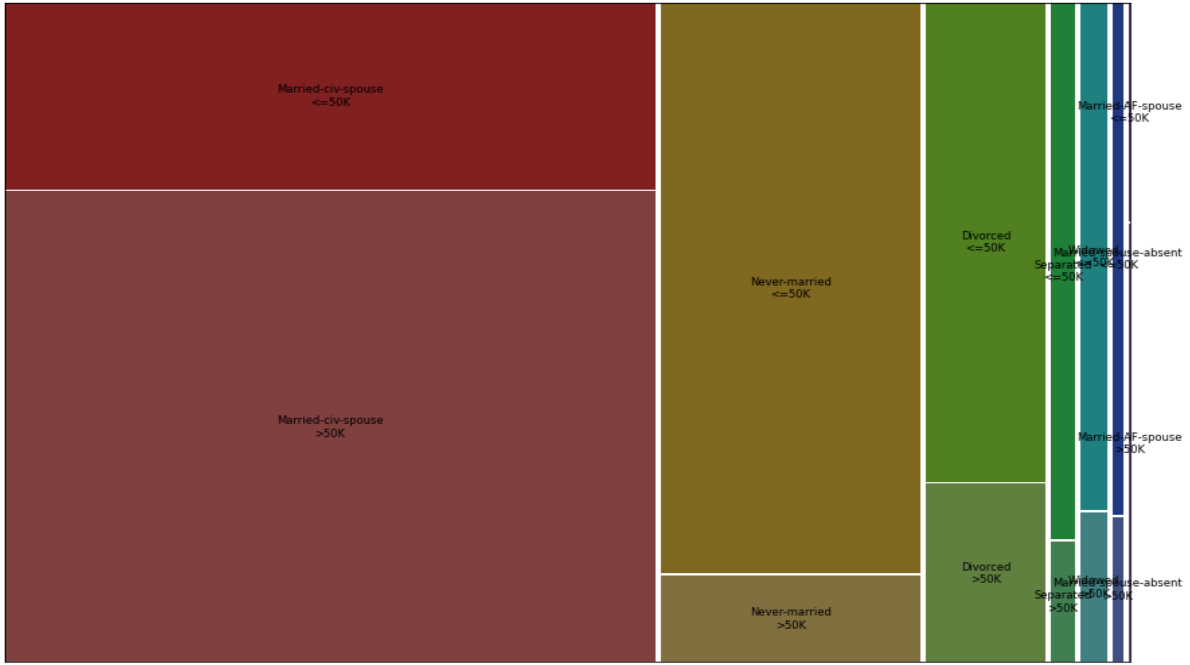
Multivariate Analysis

Marital Status and Salary

Comparing the sizes/lengths of rectangles corresponding to the values of “Married civ-spouse” and “never married”, we can say that marital status and salary range are highly skewed.

<u>Inference</u> - Individuals with the marital status of “Married civ-spouse” are more likely to earn a salary >= 50k.
Hence, XYZ corporation can use Marital Status as a feature for their prediction model.

Mosaic Plot displaying Marital-Status ,Salary-Range

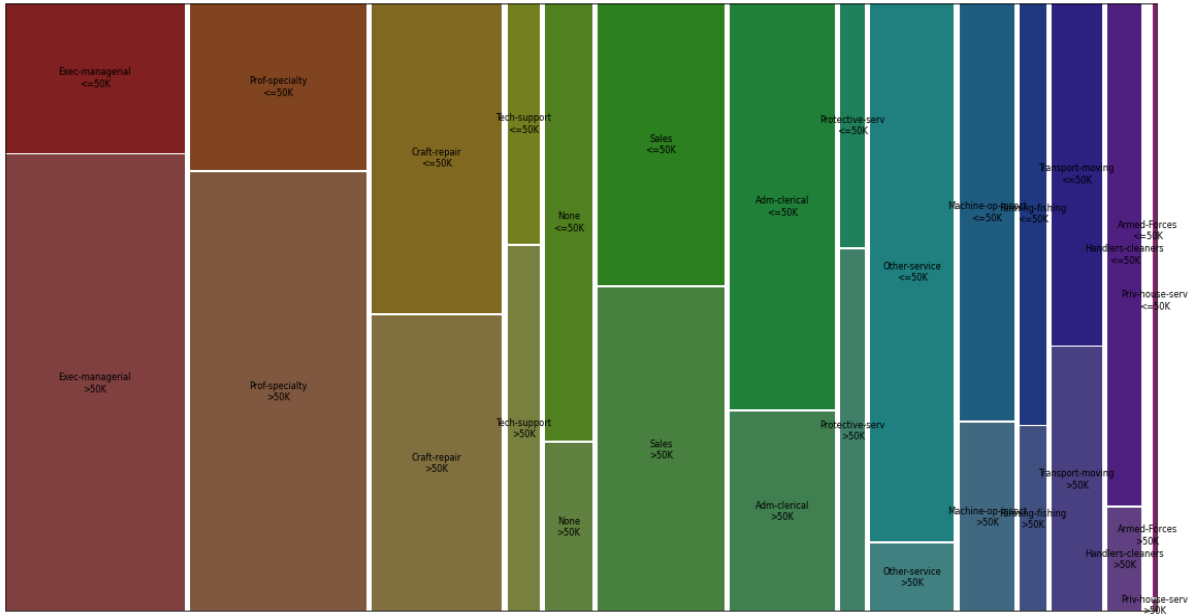


Occupation and Salary

Comparing the sizes/lengths of rectangles corresponding to the values of “Exec-managerial” and “Prof-Speciality”, we can say that Occupation and salary are skewed enough.

<u>Inference</u> - Individuals with occupations of “Exec-managerial” and “Prof-Speciality” are more likely to earn a salary >= 50k.
Hence, XYZ corporation can use Occupation as a feature for their prediction model.

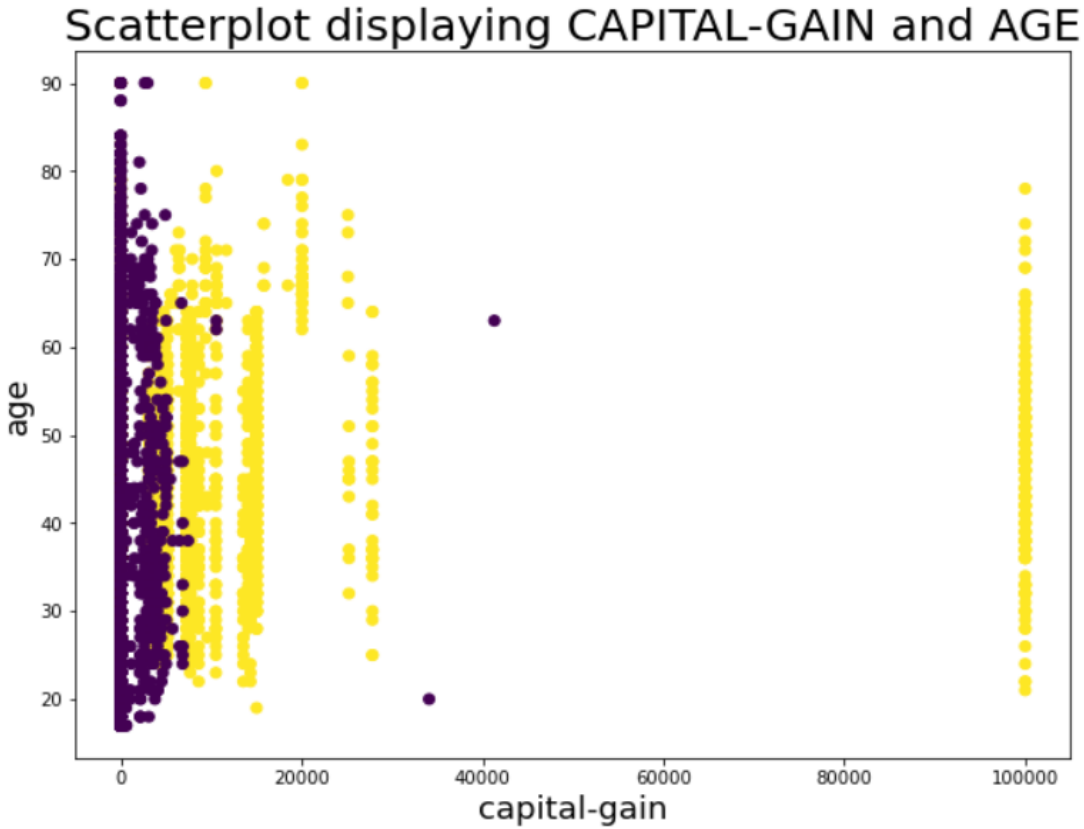
Mosaic Plot displaying Occupation ,Salary-Range



Capital Gain

There is a very clear separation between two classes of salaries, from the above scatter plot with the exception of few outliers.

<u>Inference</u> - Individuals with high "capital gain" are more likely to earn a salary above >=50k.
Hence, XYZ corporation can use Capital Gain as a feature for their prediction model.



Project Contribution

Raghu	Data Cleaning , Multivariate Data Visualizations, Report , ML analysis
Meghana	Data Exploration , Univariate Data Visualizations, Report, ML analysis
Nithin	Data Cleaning , Univariate Data Visualizations, Report , ML analysis
Rishika	Data Exploration , Multivariate Data Visualizations, Report, ML analysis
Keerthi	Data Exploration , Multivariate Data Visualizations, Report, ML analysis
Susant	Data Cleaning , Univariate Data Visualizations, Report , ML analysis

THANK YOU