CSE 578 - DATA VISUALIZATION
PREDICTION OF FACTORS INFLUENCING SALARY

Team 8

Name	ASU ID
Keerthi Reddy Ravula	1219622872
Meghana Sukhavasi	1220187449
Rishika Bathini	1220439519
Raghunath Reddy Mandapati	1219498514
Nithin Jayakar Padala	1220061427
Susant Kumar Palai	1219501478

Summary :

Our project aims to develop marketing profiles for an XYZ corporation where we operate as Data Analysts. We establish different marketing profiles for UVW college to derive the input data from the United States Census Bureau. Building marketing profiles is dependent on each individual's salary/income predicted from the features, which will allow the company to target based on their earnings. We choose the essential features which will help the prediction model to attain high accuracy.

Domain Abstraction:

Our XYZ company works on designing marketing profiles for various clients so that marketing companies can approach them individually based on their characteristics. Non-numerical data is gathered and analysed through evaluation, interviews, focus groups, and record analysis. We equate new data to old data over time while gathering data and establishing connections between new and old data.

Roles and Responsibilities :

- Team members: Keerthi Reddy Ravula, Meghana Sukhavasi, Rishika Bathini, Raghunath Reddy Mandapati, Nithin Jayakar Padala, Susant Kumar Palai.
- Stakeholders: UVW college
- Product Owners: XYZ corporation

The primary tasks involved are understanding the dataset, data cleaning, identifying features crucial to label prediction, and building machine learning models using the features to forecast everyone’s income. At first, each member of the team had to learn the dataset on their own. We had equally spread the features among the team members to evaluate them against the label (salary) and come up with their conclusions.

The task distribution is as follows:

Data Exploration	Keerthi, Rishika, Meghana
Data Cleaning	Raghunath, Susant, Nithin
Data Visualizations - Univariate	Nithin, Susant, Meghana
Data Visualizations - Multivariate	Rishika, Raghunath, Keerthi
Report Writing & PPT presentation	All team members
Machine Learning Analysis	All team members

Team Goals and Business Objective:

This project seeks to define trends in the dataset using visualizations to assess the features that influence an individual's income and present the findings to UVW executives. Following that, we'd like to build machine learning models based on the initial study that reliably forecast individual revenue. The UVW marketing team will then customize their marketing efforts to reach out to the individuals based on the application and research findings.

Assumptions:

We have made the following assumptions for various components.

- Data Accuracy: We assume that the data presented is accurate, has the correct domain. For example, if a certain age is recorded incorrectly, something like 1000 would bias the prediction model.
- Data Collection Methodologies are unbiased: Data processing should be consistent, and it should avoid colliding with any other circumstance. For example, many people are either unemployed or receiving unemployment benefits because of COVID. Data collection at such a time would be inaccurate, resulting in incorrect estimates.
- Prediction tools are working up to the mark: In our modelling, we have used several visualization tools and prediction models. We assume that these tools are accurate.
- Reliability: Data should not contradict significantly from other resources

User Stories:

The XYZ Corporation is developing personal marketing profiles to target marketing to increase UVW college enrolment. This project's team needs to create individual profiles focused on salary statistics. For instance, if most people earn less than $50000 a year, are male, and under the age of 30, tailored advertising (perhaps lower tuition, curriculum focus that leads to a higher income, etc.) can be targeted for this demography.

Hence, it is of utmost importance to determine which specific attributes have a higher contribution towards the income prediction; the accuracy of this prediction depends on these chosen features.

The XYZ corporation has collected data on 14 different attributes, and under this project, we evaluated the following user stories.

#1. Analyse if age is a relevant factor for predicting the label.

#2. Analyse if the sex of an individual is a important factor for predicting the Income.

#3. Analyse if the education level of an individual is a significant factor for predicting the label.

#4. Analyse if the marital status of an individual is a essential factor for predicting the income label

#5. Analyse if the occupation of an individual is a relevant factor for predicting the income label.

#6. Analyse if the relationship status of an individual is a meaningful factor for predicting the label.

#7. Analyse if capital gain made by an individual is a significant factor for predicting the income label.

#8. Analyse if capital loss incurred by an individual is a vital factor for predicting the income label.

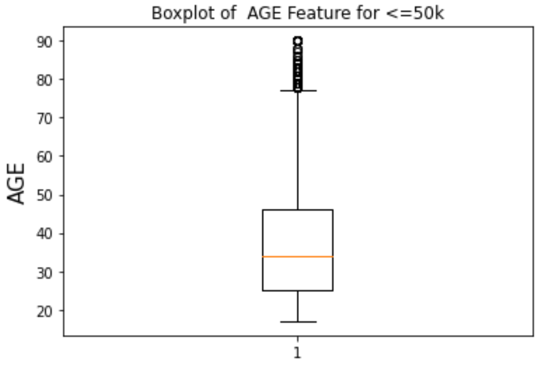
#9. Analyse if the number of work hours per week is a relevant factor for predicting the income label.

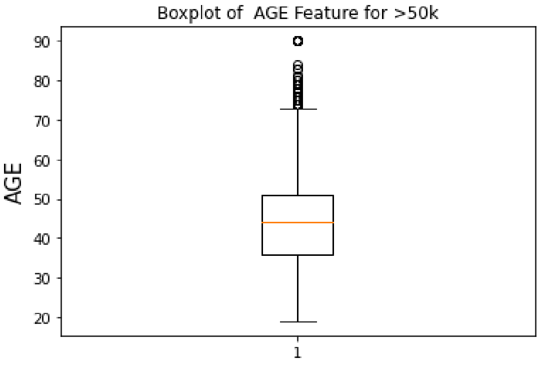
#11. Analyse if the native country of an individual is an essential factor for predicting the income label.

<u>Visualizations</u>:

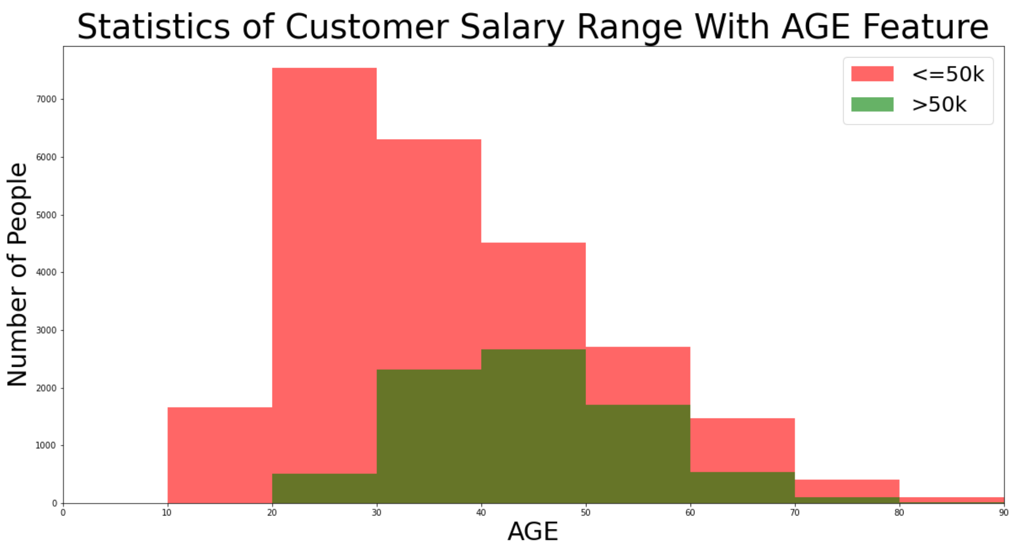
Univariate Analysis:

Feature: Age





The above boxplot shows that there are outliers in the date for age feature.

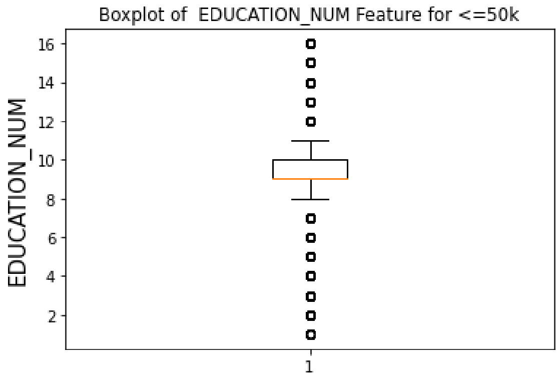


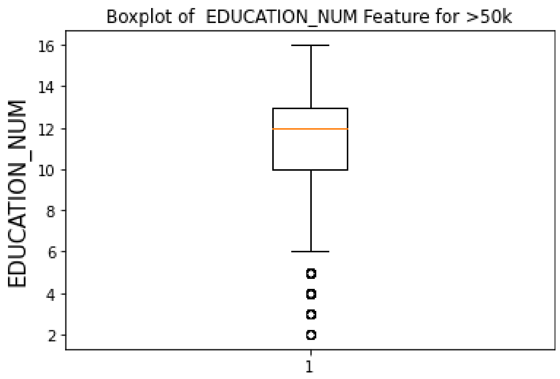
Age is continuous numerical data, so a histogram is an excellent option to visualize the trend. We need to compare the data whose salary is >50k and <=50k; thus, it is easier to reach a stacked bar chart. Overlapping histograms allow viewers to compare data quickly, and besides, they work well with large ranges of information.

Inference - 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.

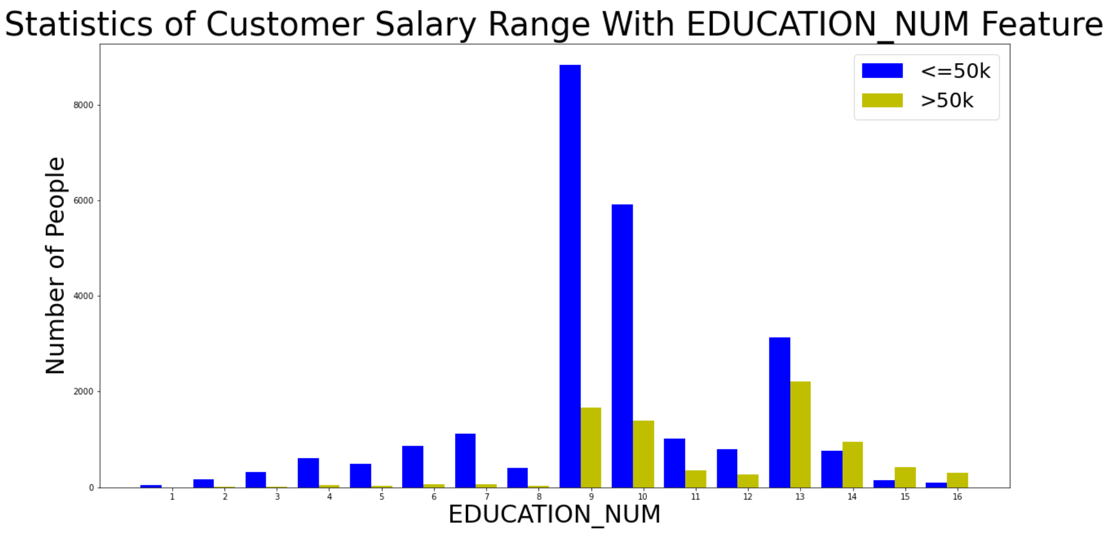
Hence, XYZ corporation can use ‘age’ as a feature for their prediction model.

Feature: Education Num





The above boxplot shows that there are outliers in the data for the Education Num feature.

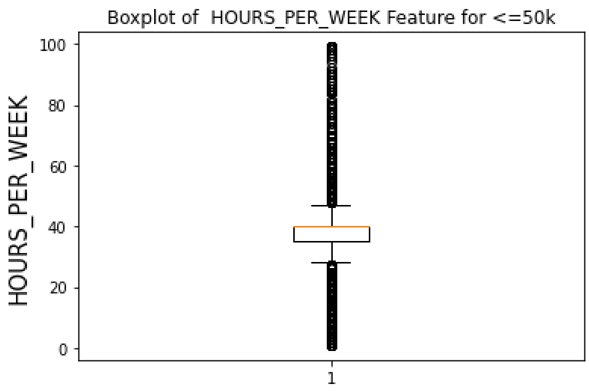


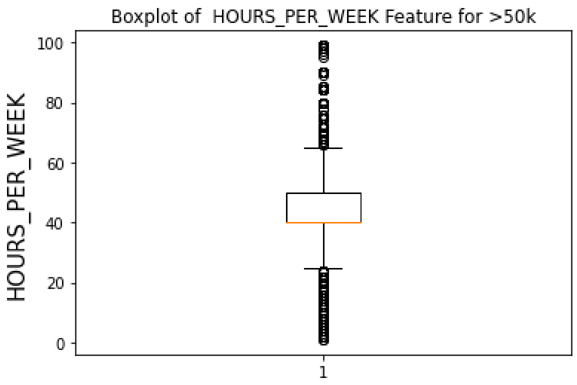
Education Num is continuous numerical data with an extensive range. A bar graph with two separate bins with salary >50k and <=50k visualizes key values at a glance, which allows a visual check of the correctness and reasonableness of estimates.

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

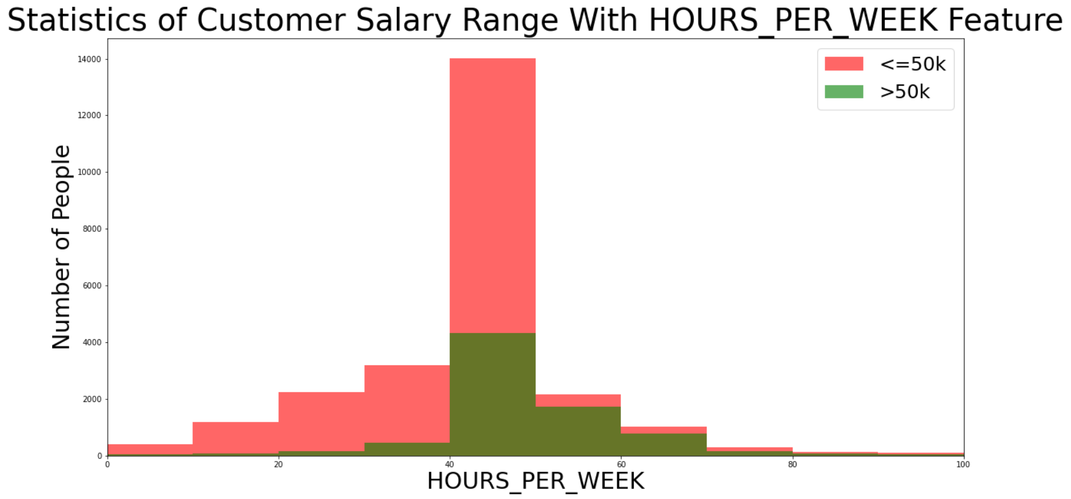
Hence, XYZ corporation can use Education Num as a feature for their prediction model.

Feature: Hours Per Week





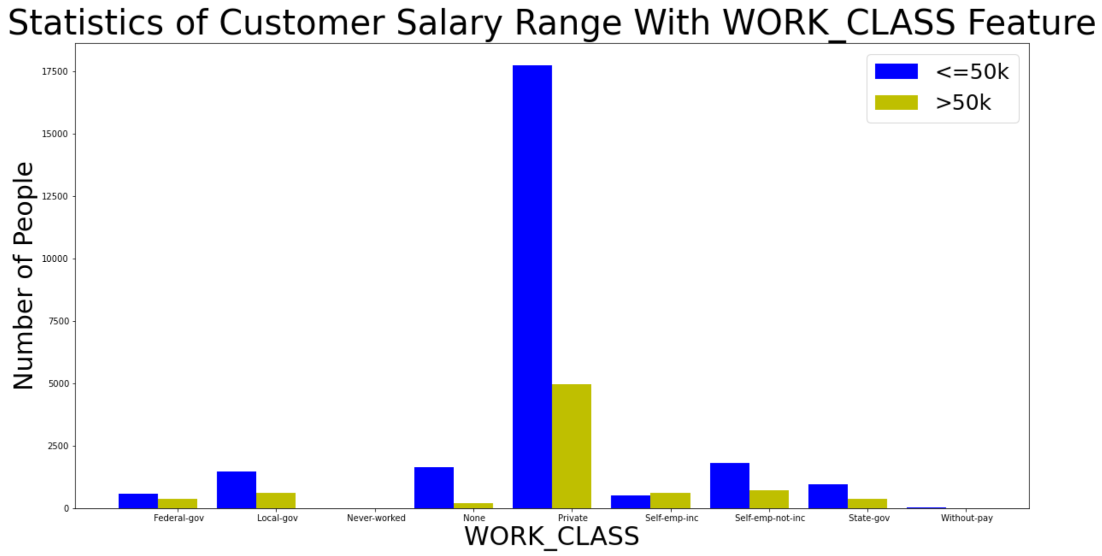
From the above boxplot, most of the outliers are in the region other than the range of hours_per_week (40, 50).



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.

Inference - 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.

Feature: Work Class

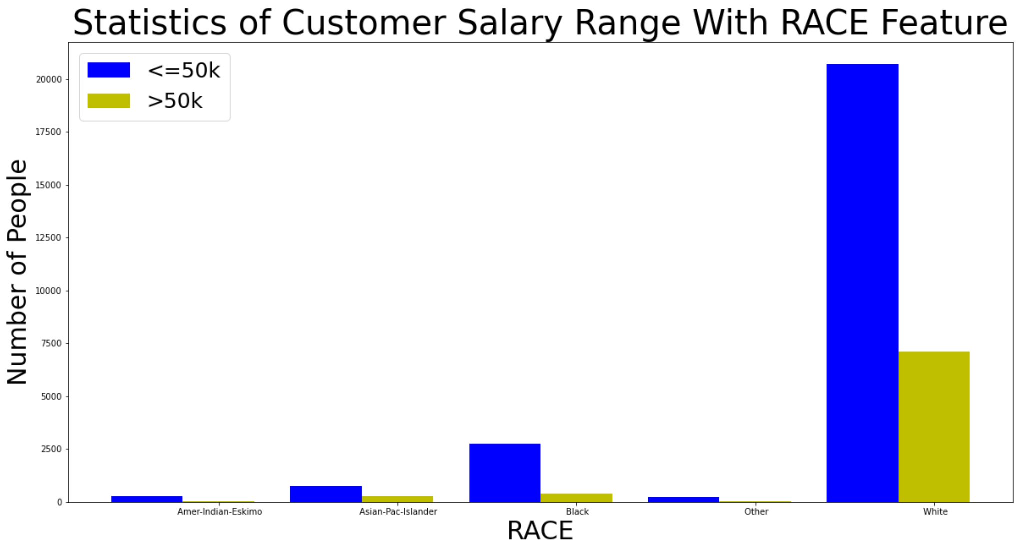


Work Class is a discrete categorical data. A bar graph with two different bins for salaries >50k and <=50k visualizes central values at a single glance.

Inference - 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.

Feature*:* Race

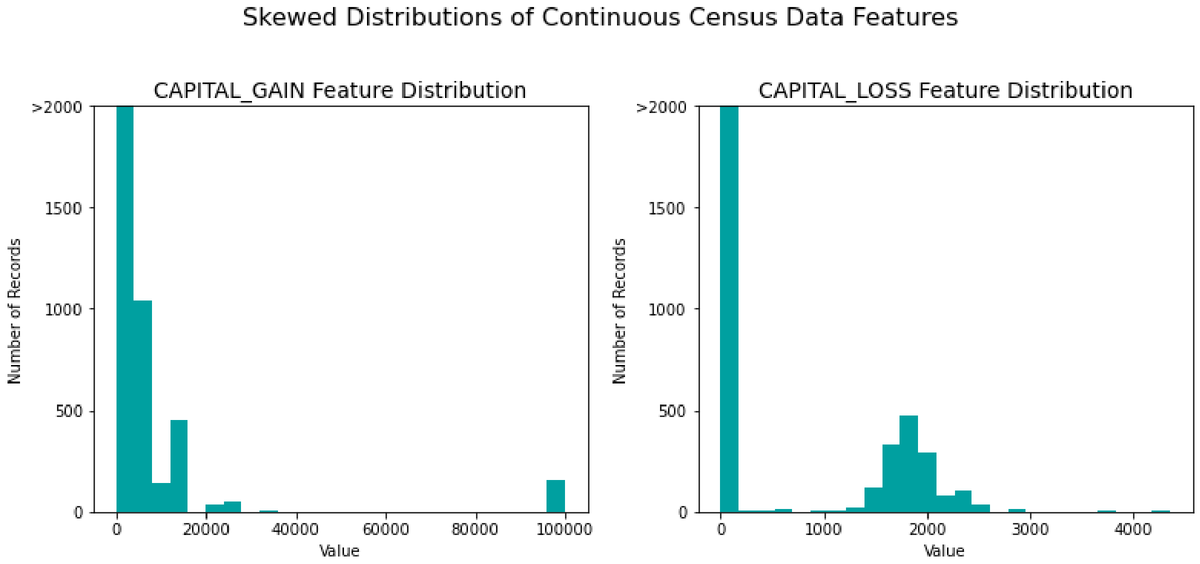


The race is discrete categorical data. A bar graph with two different bars for salaries >50k and <=50k visualizes central values at a glance, allowing a visual check of the correctness and reasonableness of estimates. By looking at the above histogram, the data is left skewed at the race "White"

Inference - 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.

Feature*:* Capital Loss, Capital Gain



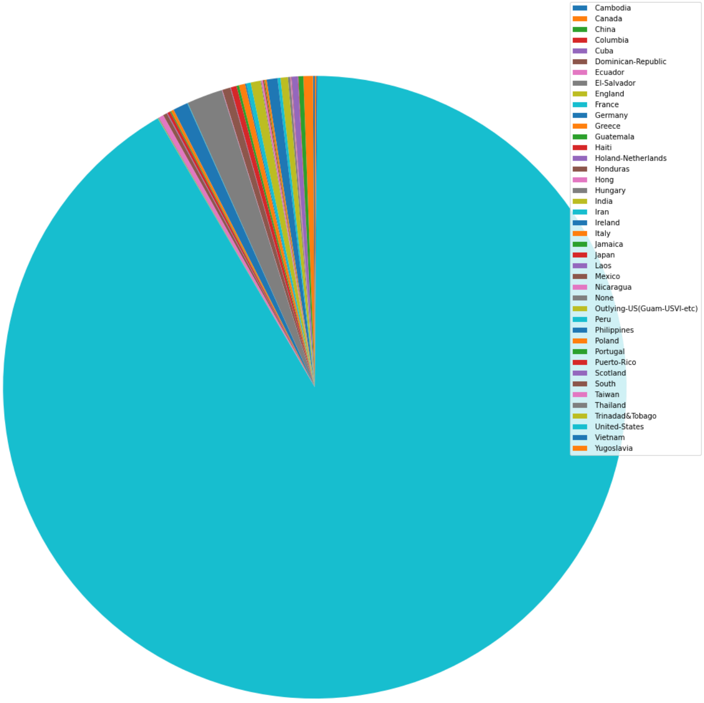
We have used Histograms for Capital loss, and Capital Gain features are continuous numerical variables.

Inference - There is a strong indication that people with capital gains between 1k - 20k are likely to earn >50k.

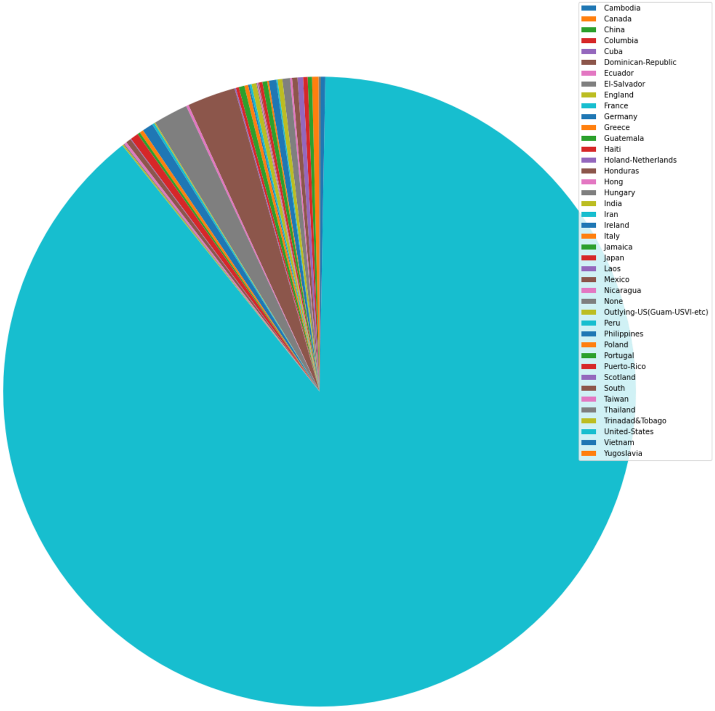
Hence, XYZ corporation can use Capital Loss, Capital Gain as a feature for their prediction model

Feature: Native Country

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



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

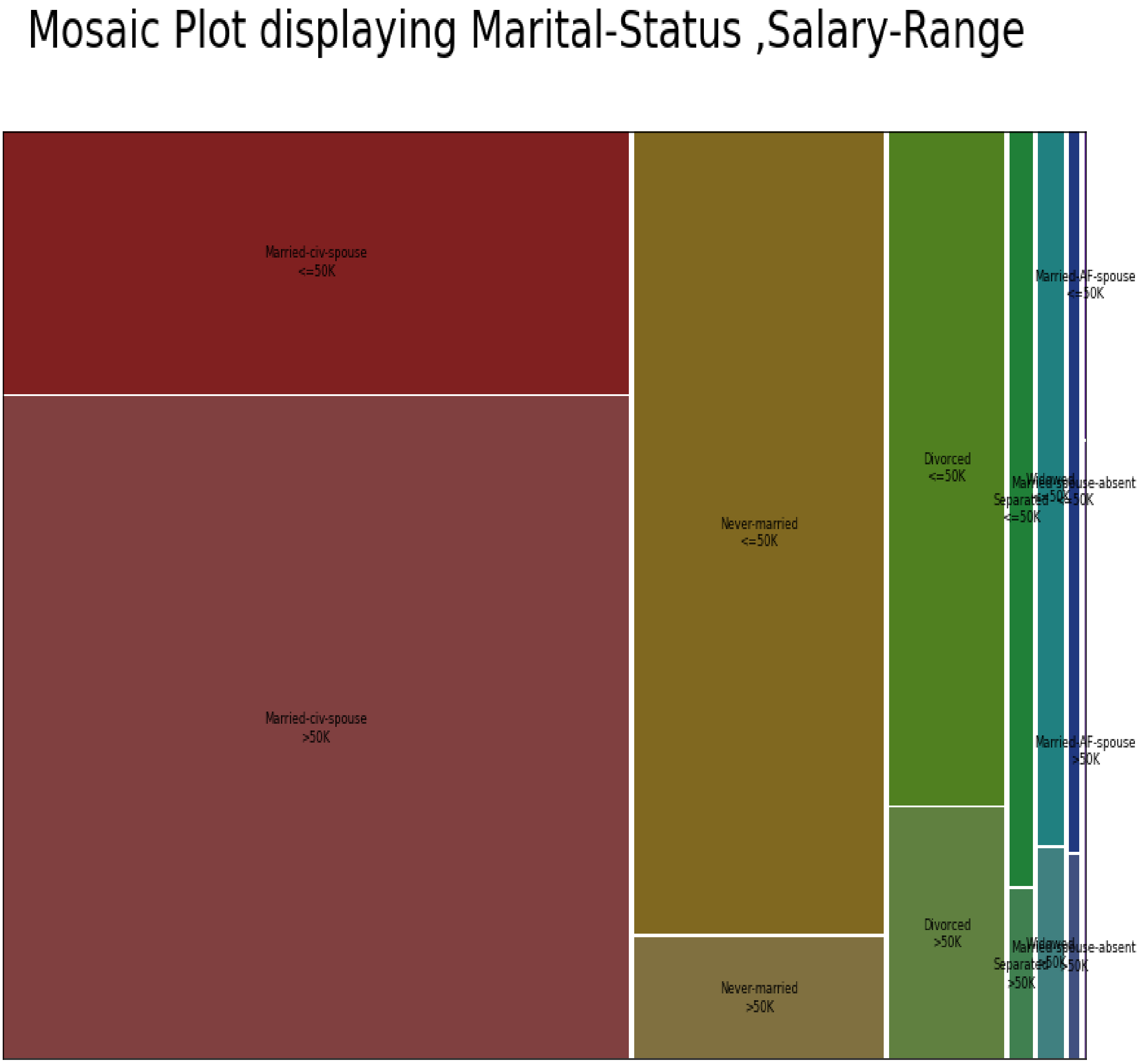


The Native Country feature is discrete categorical data with many categories. Pie charts are good at relative proportions of multiple data classes and summarize an extensive data set in the visual form; thus, we have used Pie charts for Native Country.

Inference - There are many individuals from the United States whose salary > 50k.
Hence, XYZ corporation can use Native Country as a feature for their prediction model.

Multivariate Analysis:

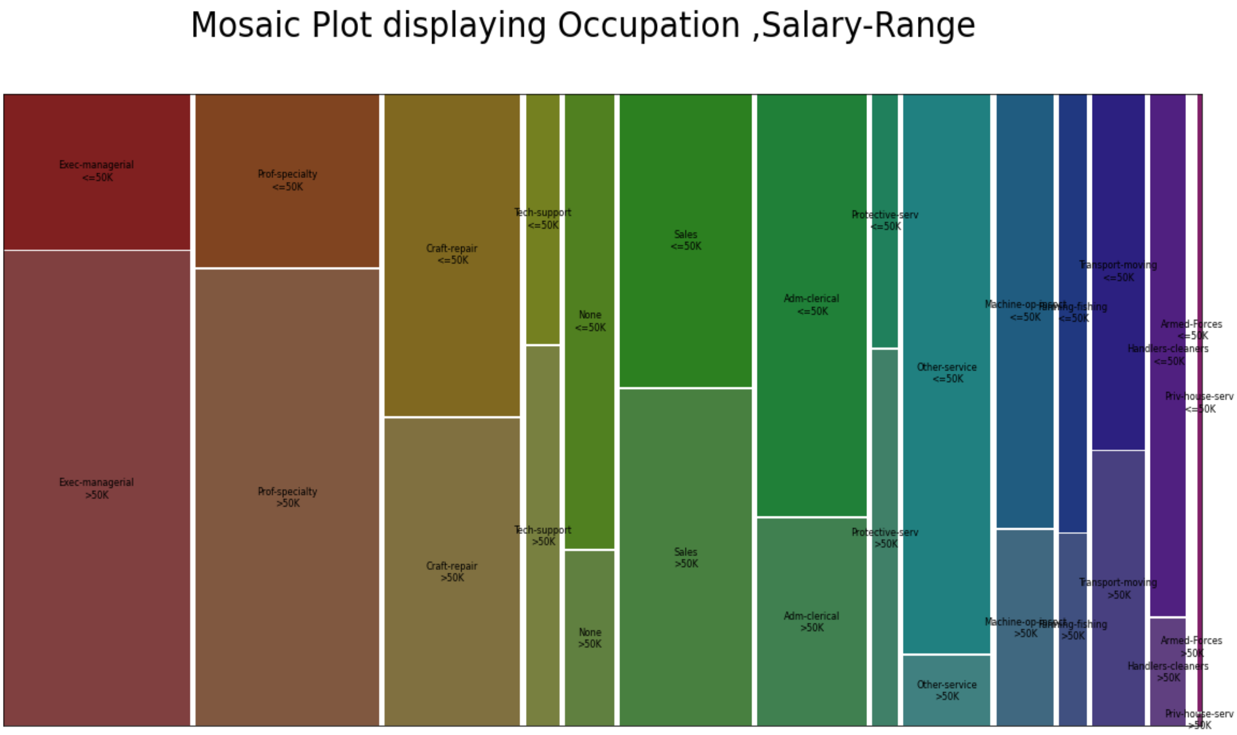
Feature: Marital Status and salary



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.

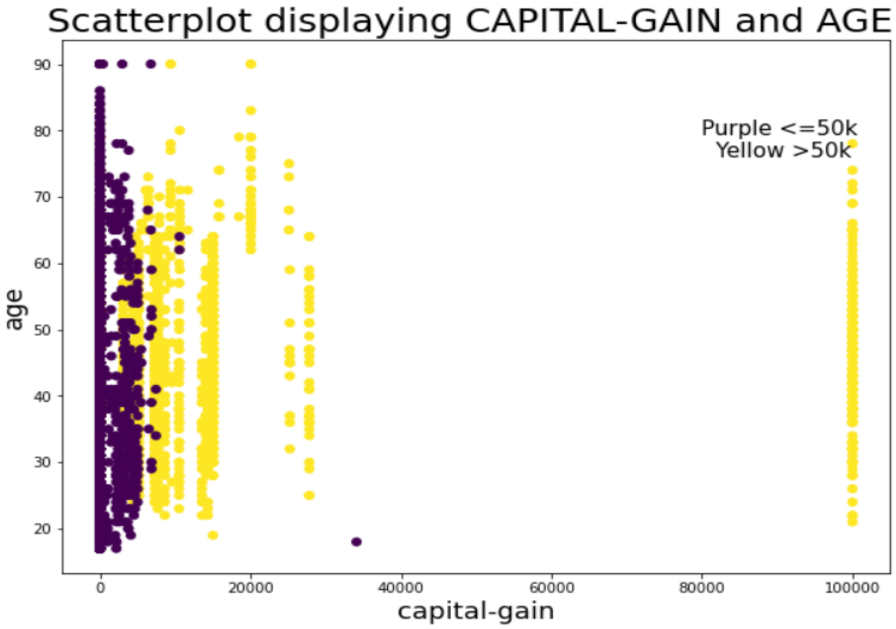
Inference - Individuals with the marital status of “Married civ-spouse” are more likely to earn a salary >= 50k.
Hence, XYZ corporation can use Marital Status as a feature for their prediction model.

Feature: *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.
Inference - 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.

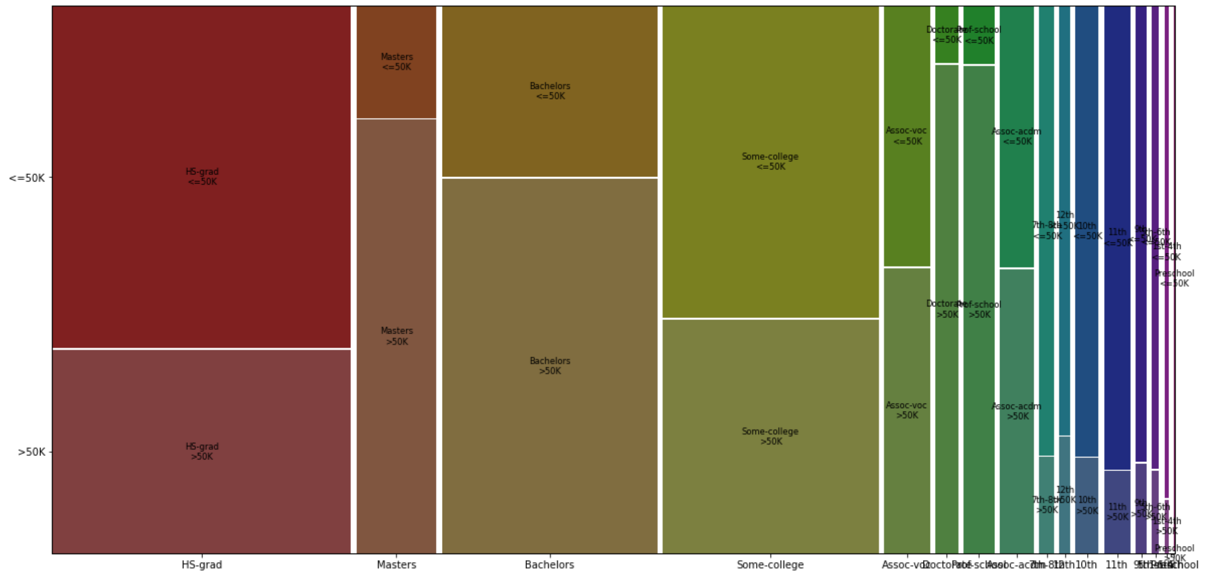
Feature: *Capital-gain*



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

Inference - 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.

Feature: *Education* (Redundant)



Inference – We can clearly say that the data is evenly distributed in this mosaic plot. Hence, XYZ corporation cannot use Capital Gain as a feature for their prediction model.

<u>Tools:</u>

- Python, Jupyter Notebook, Matplotlib, Numpy, Pandas, Sklearn.

<u>Machine Learning Analysis</u>:

We have attached the Jupyter Notebook of Machine Learning Analysis for reference.
Involved Steps:

1. **Feature Engineering** - We are leaving all the numerical attributes as they are. From our initial data exploratory analysis, we are arriving at the distinguishing factor of each category. For each such categorical data, a numerical number is assigned based on the arrived distinguishing factor.
2. **Normalization of Data** - We are scaling each attribute to some value between 0 and 1 to ensure equal importance is given to ML algorithms.
3. **Division of Data** - In the data provided to us, --- samples are given to use for testing. The remaining data is being divided in the ratio of 80:20, 80% would be for the training set, and 20% would be for the validation set.
4. **Training ML Models** - We have used the following Machine Learning methods for training purposes. Gaussian Naïve Bayes, Decision Tree, Ensemble Methods (Random Forest), SVM, Logistic regression, SGDC.

5. **Tuning of Hyperparameter** - For each of the above algorithms, the hyperparameters are altered to obtain the most relevant results.
6. **Performance Evaluation** - Evaluation is done by trained models using different measures like F1 score, accuracy, precision, and recall.

Questions :

Q: How to get rid of the noisy data as we see a lot of empty spaces and question marks in the given dataset.

Here, we follow the process of data cleansing where noisy data could be avoided. After careful observation of data, we found out there was a lot of "?" which we did replace by spaces using data frames.

Q: How to get the names of the 15 attributes given in the input dataset.

We have given columns specific names to our 15 attributes based on the input data in our dataset.

Q: What features to use and how to calculate which feature outperforms the other feature?

There are 14 features in our input data set. Since all features do not contribute towards class prediction. We have individually examined each of the features using data visualization tools to understand their distribution and rank their importance. We have selected those features that have more noticeable differences in the patterns for the two classes for a higher rank.

Q: Which visualizations to use to strongly depict the data?

We have used univariate visualizations and multivariate visualizations depending on our features. We have plotted 10 Univariate visualizations using Salary and another feature against each other.

Q: Out of all the machine learning algorithms, which one to be used to give us accurate results?

We have used the following ML algorithms for our classification - Gaussian Naive Bayes, Decision Tree, Ensemble Methods, SVM, Logistic regression, SGDC. We have implemented each algorithm and with accurate results, we have decided to use these algorithms against the test data.

Not Doing:

- We are planning to extend the prediction of our class labels to different salary ranges, i.e., predicting the salary of different ranges and of different ages.
- We have selected only a few Machine Learning algorithms in implementing our current prediction model; we might use other different classification algorithms for different data sets in further implementation.