Contents

[Introduction: Overview of dataset, EDA and Classifiers 3](#_Toc192456268)

[Dataset Overview: 3](#_Toc192456269)

[EDA: 3](#_Toc192456270)

[Correlation: 4](#_Toc192456271)

[Distributions: 5](#_Toc192456272)

[Income vs Age 5](#_Toc192456273)

[Income vs Captial-Gain 6](#_Toc192456274)

[Income vs Captial-Loss 6](#_Toc192456275)

[Occupation Distribution by Income: 7](#_Toc192456276)

[Marital Status Distribution by Income: 8](#_Toc192456277)

[Gender Distribution by income: 9](#_Toc192456278)

[Relationship Status Distribution by Income: 10](#_Toc192456279)

[Country-wise Distribution 12](#_Toc192456280)

[Classifiers: 12](#_Toc192456281)

[Data Preprocessing: 13](#_Toc192456282)

[Imputation Methods: 13](#_Toc192456283)

[Mean Imputation: 13](#_Toc192456284)

[Mode Imputation: 13](#_Toc192456285)

[Linear Regression Imputation: 13](#_Toc192456286)

[Scaling Methods and its justification: 13](#_Toc192456287)

[Log Scaling: 13](#_Toc192456288)

[Robust Scaling: 13](#_Toc192456289)

[Label Encoding: 14](#_Toc192456290)

[Model Comparisons and discussion: 14](#_Toc192456291)

[**Pipeline Model P1: Linear Regression Imputation + Log Scaling Logistic Regression (regularized L1)** 14](#_Toc192456292)

[**P2: Linear Regression Imputation + Robust Scaling +Logistic Regression** 15](#_Toc192456293)

[**P3: Mean Imputation + Log Scaling Logistic Regression (regularized L2)** 16](#_Toc192456294)

[**P4: Mean Imputation + Robust Scaling Logistic Regression** 17](#_Toc192456295)

[P1: Linear Regression Imputation + Log Scaling SVM (regularized L1) 18](#_Toc192456296)

[P2: Linear Regression Imputation + Robust Scaling SVM 18](#_Toc192456297)

[P3: Mean Imputation + Log Scaling SVM (regularized L2) 19](#_Toc192456298)

[P4: Mean Imputation + Robust Scaling SVM 20](#_Toc192456299)

[Discussion on how different preprocessing techniques influenced model performance. 21](#_Toc192456300)

[Final Observations: 21](#_Toc192456301)

[Identify the best-performing model 21](#_Toc192456302)

[Explain the impact of imputation and scaling 22](#_Toc192456303)

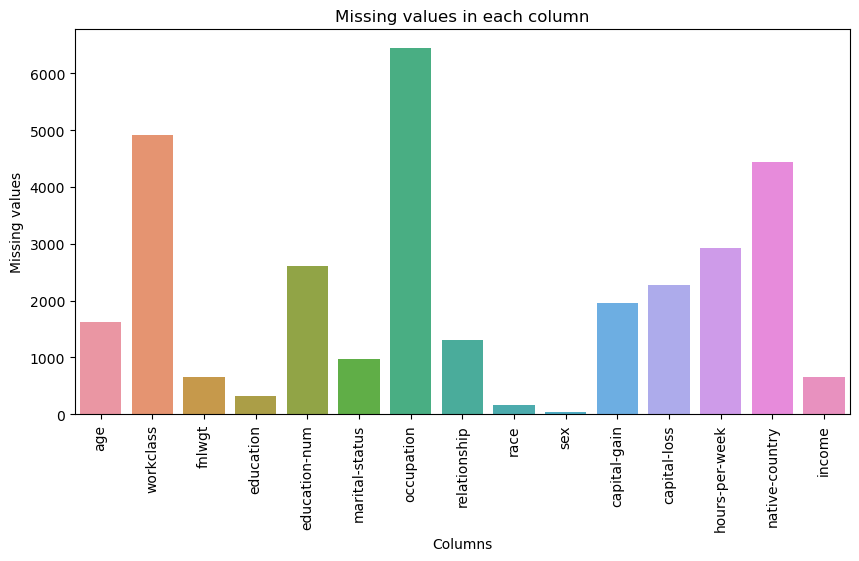
[Recommendations on when to use different preprocessing approaches. 22](#_Toc192456304)

# Introduction: Overview of dataset, EDA and Classifiers

## Dataset Overview:

In this assignment, we are required to use the dataset named “adult.csv”, and contains the information of age,workclass,education,occupation,race,sex on the basis of the income (which is they earn more than or less than $50,000 per year ).

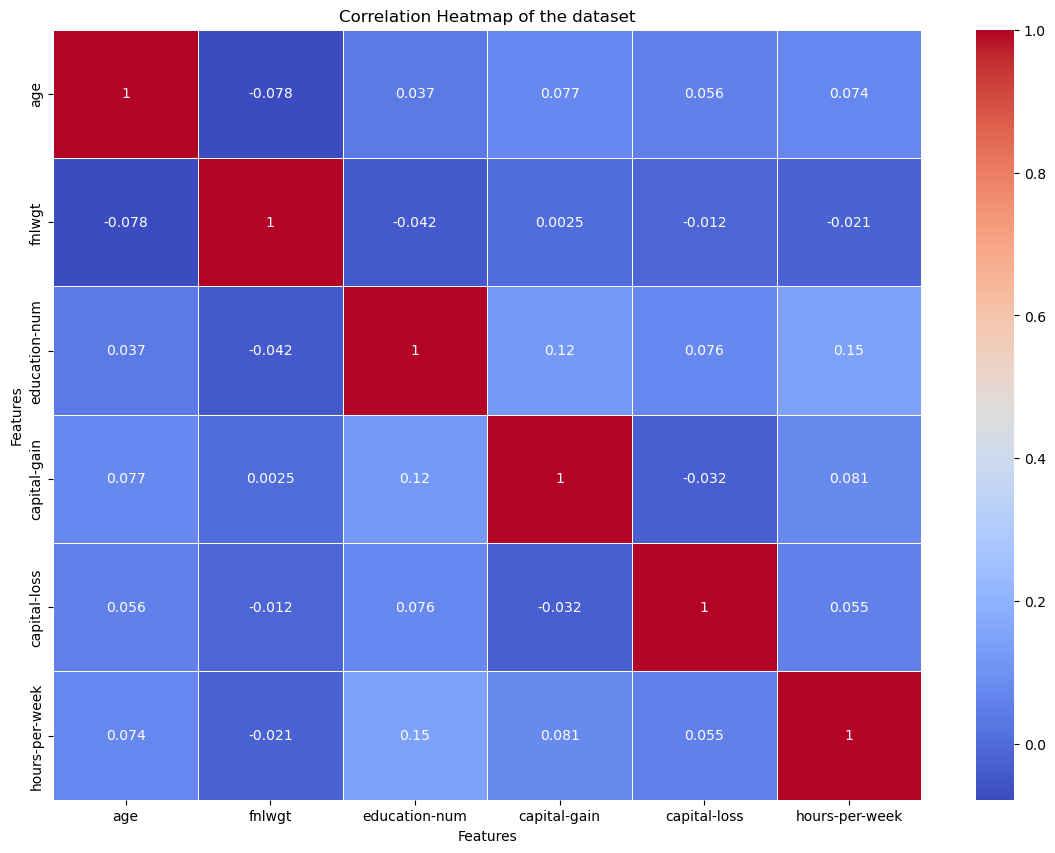
## EDA:

The issue with this dataset is that it contains question marks, spaces and empty entries which leads to incorrect prediction of the models.

By this graph, we see that the empty entries in each column of the dataset. Hence, major missing entries are “Work class”, “ Occupation ”, and “ native-country”.

Therefore, we fixed these entries by replacing them by NaN values as it is unknown to us.

## Correlation:



By looking at the correlation heatmap (it shows different correlations between each column of the dataset), we see that the highest correlation of this dataset is 0.15 between education\_num and hours\_per\_week. Therefore, it indicates that higher education level tends to work more hours.

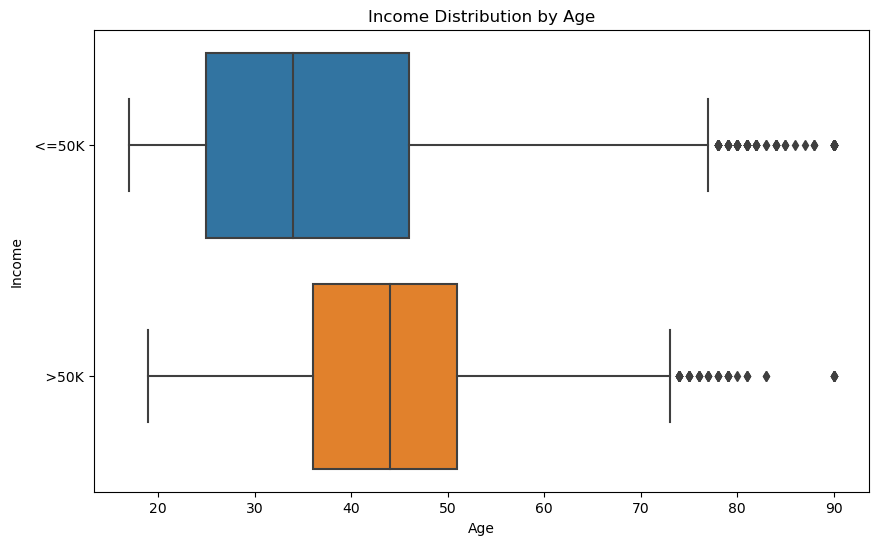
On the other hand, we see that the lowest correlation of this dataset is -0.078 between age and fnlwgt, showing that there is no almost no correlation between in these 2 features.

Furthermore, capital gain and education\_num shows the positive correlation of 0.12, indicating that higher education levels are more likely to have more capital gains.

This suggests that a combination of multiple features is necessary to build an effective predictive model.

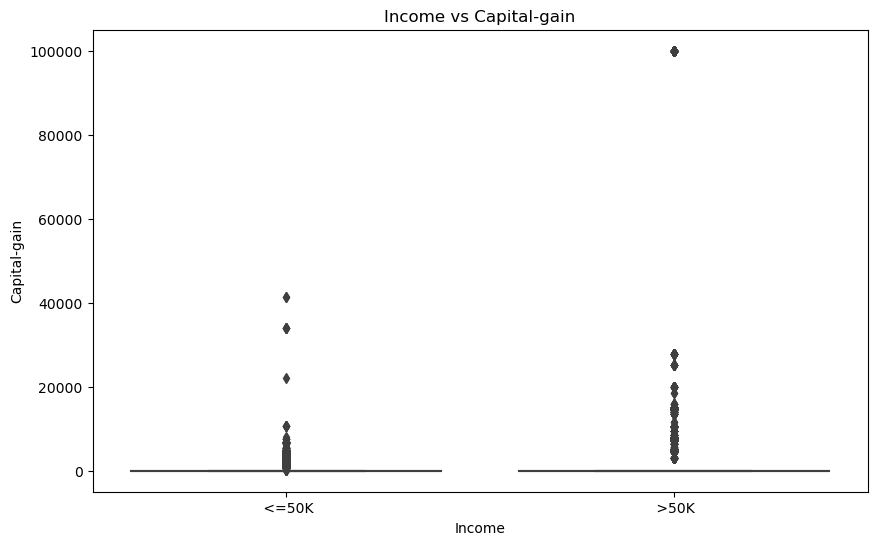
## Distributions:

### Income vs Age



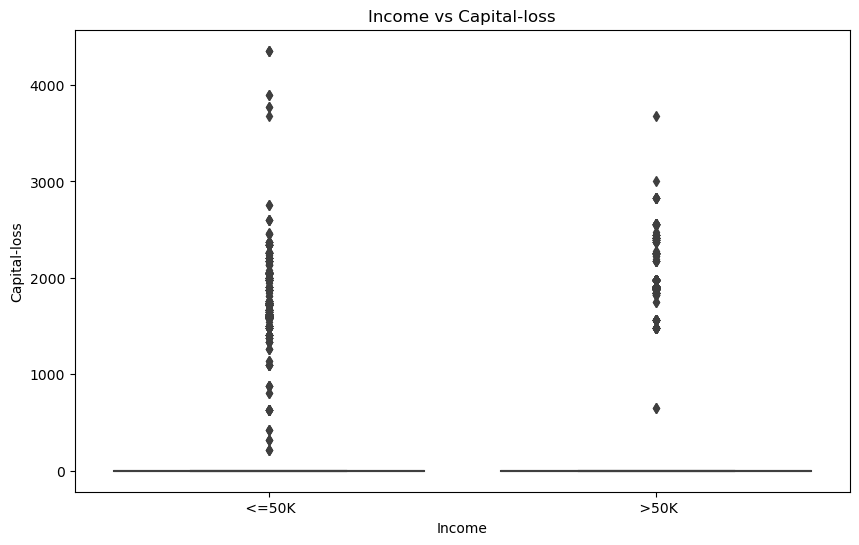
From this this box and whisker, we see the incomes varies with the age. Any person whose age is more than 40 are earning more than 50k as compared to other ages who are

### Income vs Captial-Gain



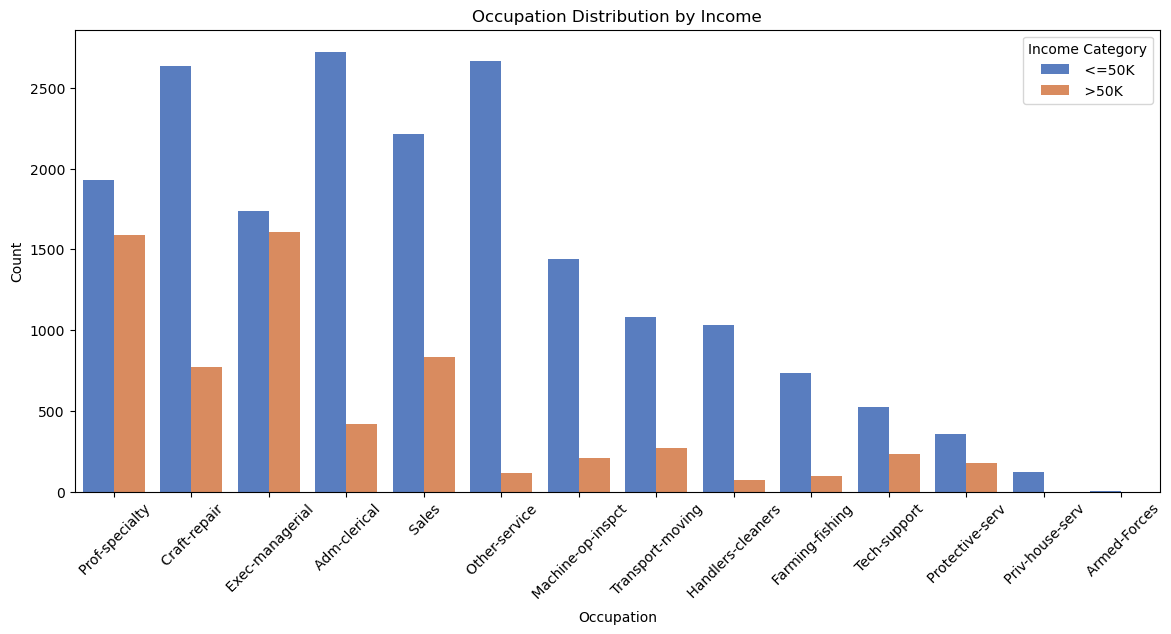
From this graph, we see that the population whose earning is less than 50k has low capital gain as compared to the population who is making more than 50k has higher capital gain

### Income vs Captial-Loss



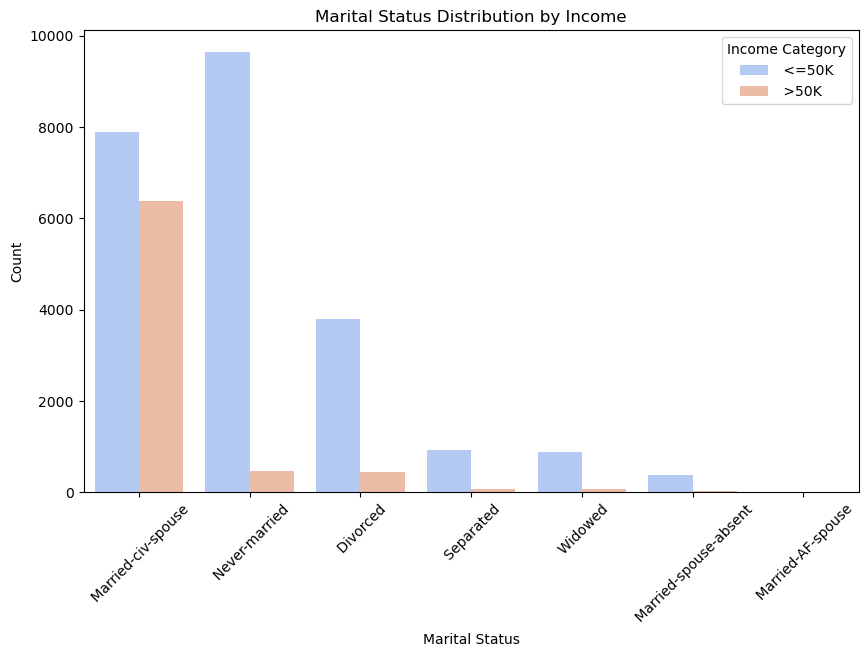
From this graph, it shows that both >50k and <=50k have a very similar capital loss and indicates that capital loss is not much of a classifier for income.

### Occupation Distribution by Income:



From this graph, we see that different occupations have different incomes. By looking at the Craft-repair, adm-clerical and other-services have the highest income in " <=50k " category. Meanwhile, the Pro-specialty, Exec-managerial and Sales have the highest income in "> 50k " category. Furthermore, the Armed-Forces are zero in terms of both income category.

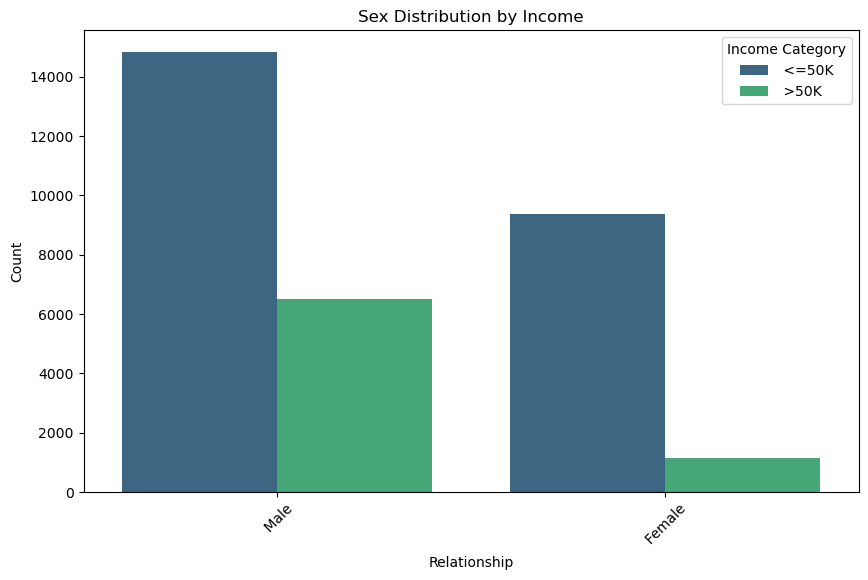
### Marital Status Distribution by Income:



From this graph, we see that highest count for the income of " <=50k " is Never-Married category. On the other hand, Married-civ-spouse category has the highest count for the income of "> 50k".

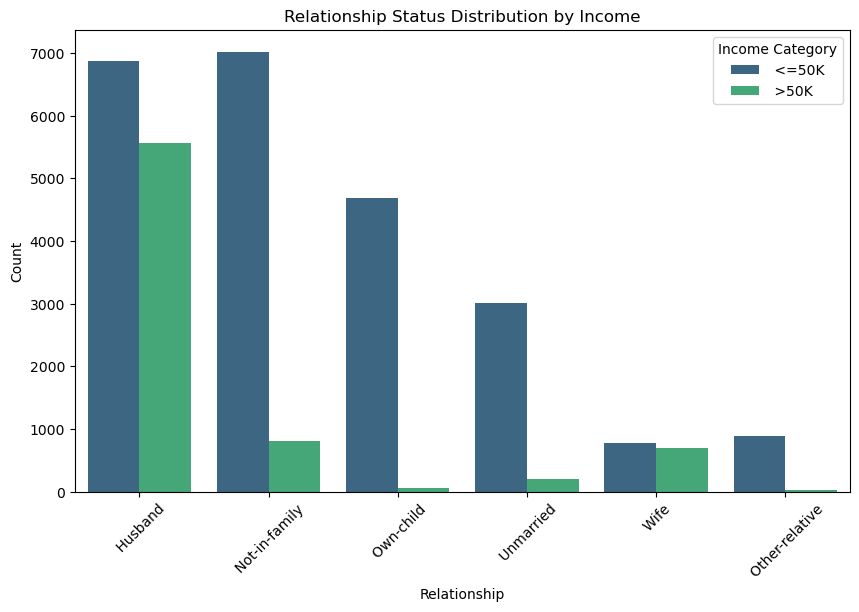
Furthermore, this category makes a good income for both of its income category (<=50k and >50k)

### Gender Distribution by income:

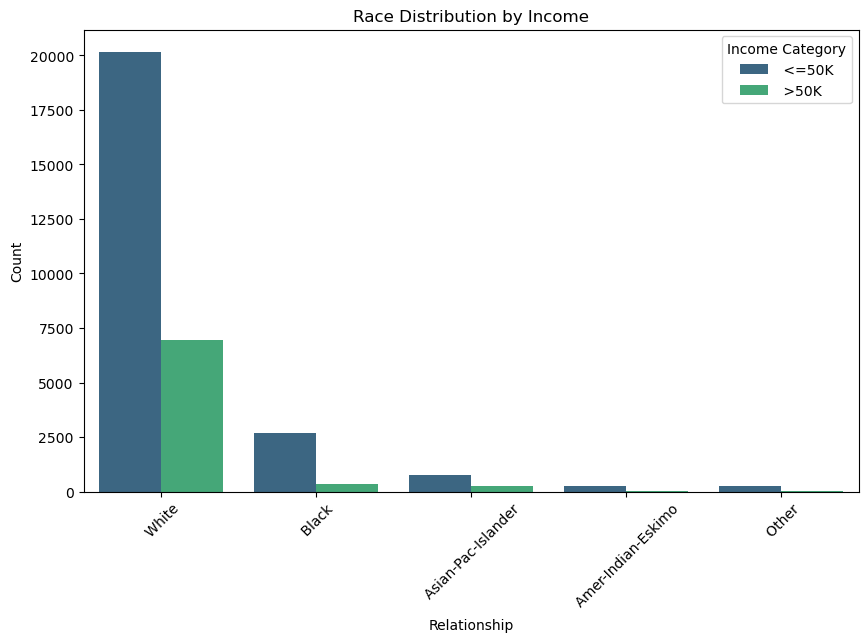


From this graph, Males form the largest group, with many earning <=50K, but a significant number also earn >50K. Meanwhile, Females have a relatively smaller distribution compared to Males, with a good portion earning <=50K and a smaller portion earning >50K. This indicates a significant income disparity between the two genders, showing that Males have higher income in both categories.

### Relationship Status Distribution by Income:

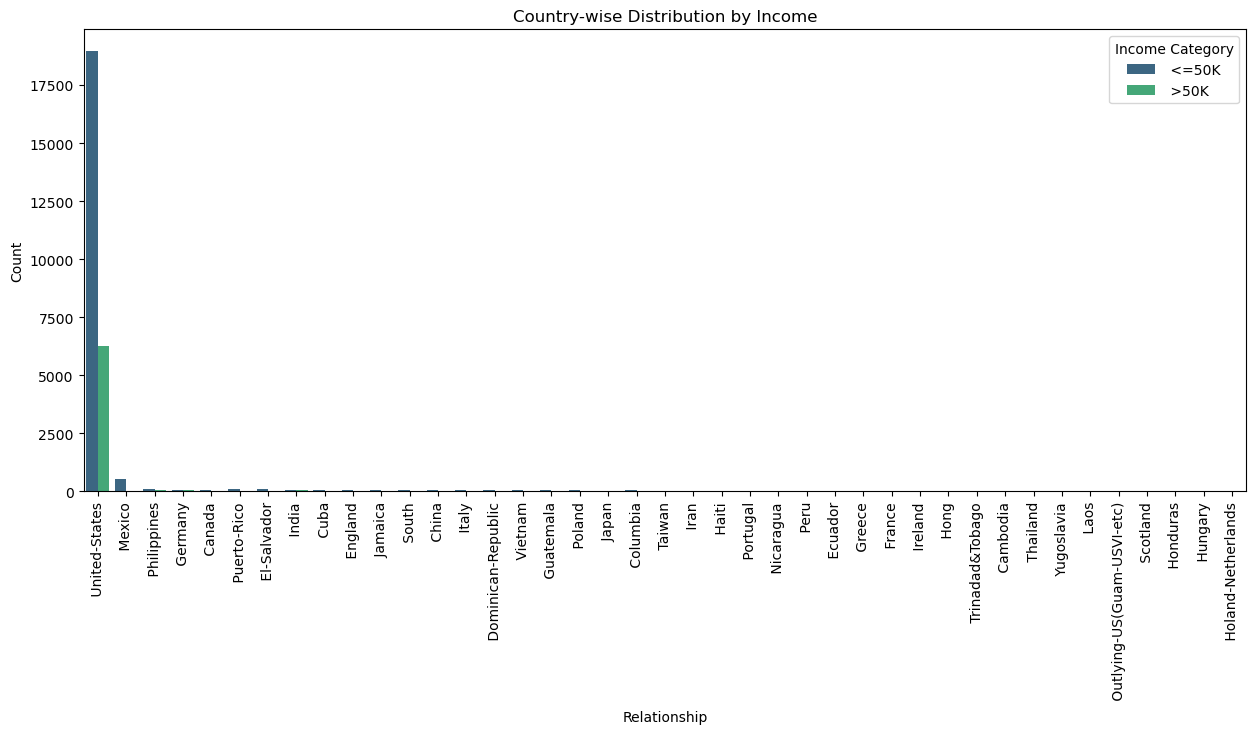


From this graph, Husbands form the largest group, with many earning ≤50K, but a significant number also earn >50K. Meanwhile, those who are not-in-family category are making the highest income of "<=50k". Furthermore, Wives have a relatively balanced distribution compared to other groups, with a good portion earning >50K.



From this graph, Whites form the largest group, have with many earnings less than 50K, but a significant number also earn more than 50K. Meanwhile, Blacks have a relatively smaller distribution as compared to Whites, with a good portion earning ≤50K and a smaller portion earning >50K. Other racial groups such as Asian-Pac-Islander, Amer-Indian-Eskimo, and Others have even smaller representations, with most individuals earning ≤50K.

### Country-wise Distribution



From the Graph, we see that the “United States” has the highest distribution than all other countries in the world. As it has significant number of earnings in respective to both more than and less than 50k. Meanwhile, Mexico has smaller distribution with most individuals having less than 50k and small portion are earning more than 50k. Other countries such as Canada, Germany, Philippines, Puerto-Rico and El-Salvador are earning less than 50k.

## Classifiers:

As we can see in the dataset, the most prominent classifiers are Income which has two values >50k and <=50k, the best feature for classification is country, education, gender and race. Meanwhile, the worst are capital gain, capital loss and fnglwt as we can see the diagrams.

# Data Preprocessing:

## Imputation Methods:

As there were missing values in the numerical columns of dataset, we are using three approaches:

* Mean Imputation
* Mode Imputation
* Linear Regression Imputation

### Mean Imputation:

In this approach, we calculate the mean value of each numerical column and replaced the missing entries with those mean values of their respective columns. This approach ensures that the overall distribution remains unchanged but may not optimal.

### Mode Imputation:

In this approach, we calculate the highest frequency of each numerical column and replaced the missing entries with those values of their respective columns. This approach ensures consistency but it may introduce bias.

### Linear Regression Imputation:

In this approach, we predict the missing values of each numerical column by using the regression models. This approach ensures consistency but it may introduce bias.

## Scaling Methods and its justification:

There are two methods for scaling approach that we used in this assignment

* Log Scaling
* Robust Scaling

### Log Scaling:

In this approach, we apply the log scaling to the numerical features to address skewness and impact of the outliers by transforming the data logarithmically. Therefore, improving the data for modelling.

### Robust Scaling:

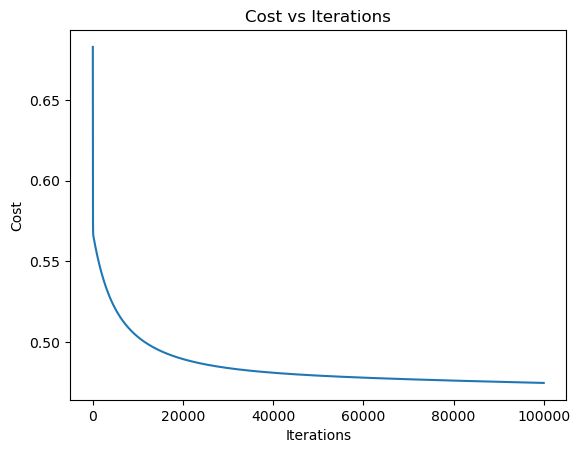
In this approach, we apply the scaling by standardizing the data using the median and interquartile range, making it effective in handling the outliers otherwise it will produce noise in the model and distort the performance

## Label Encoding:

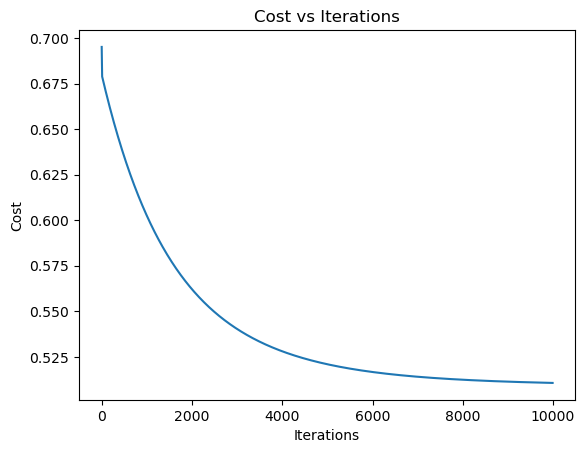
In the Label Encoding, we convert the categorical variables into numerical values, where each unique category is assigned with a corresponding integer. This approach allows machine learning to process the categorical variables efficiently.

# Model Comparisons and discussion:

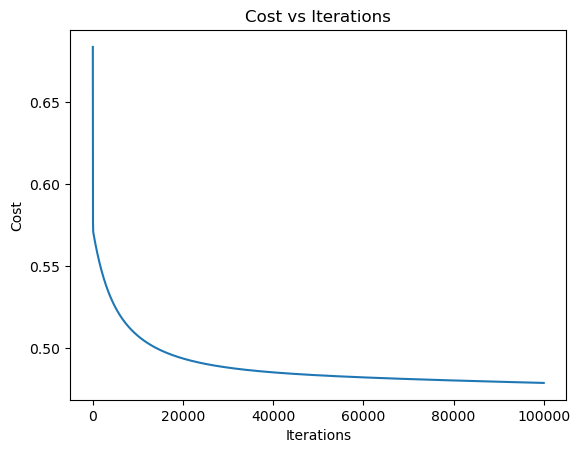
## **Pipeline Model P1: Linear Regression Imputation + Log Scaling Logistic Regression (regularized L1)**



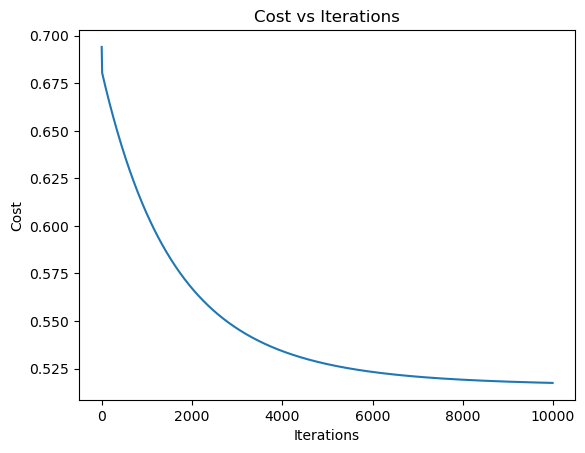
## **P2: Linear Regression Imputation + Robust Scaling +Logistic Regression**



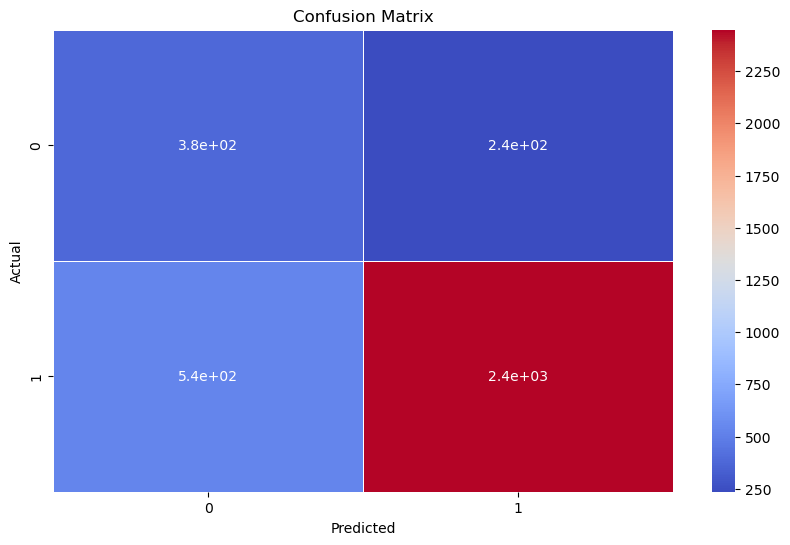
## **P3: Mean Imputation + Log Scaling Logistic Regression (regularized L2)**



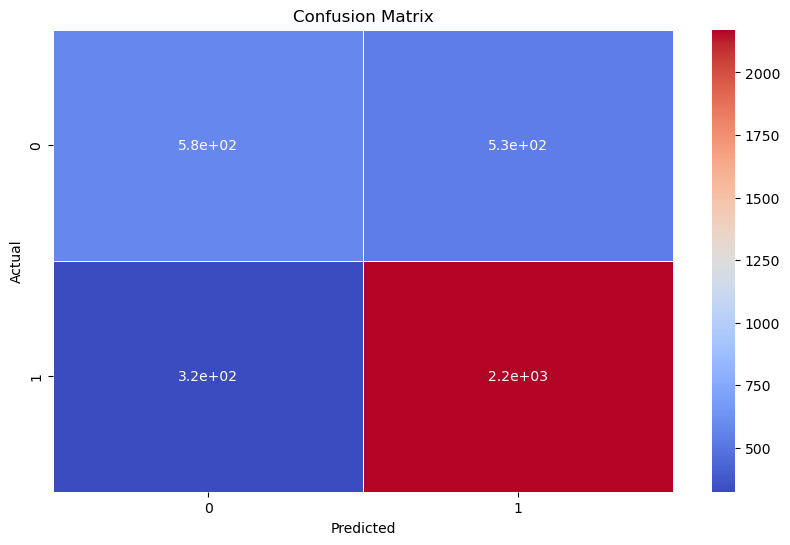
## **P4: Mean Imputation + Robust Scaling Logistic Regression**



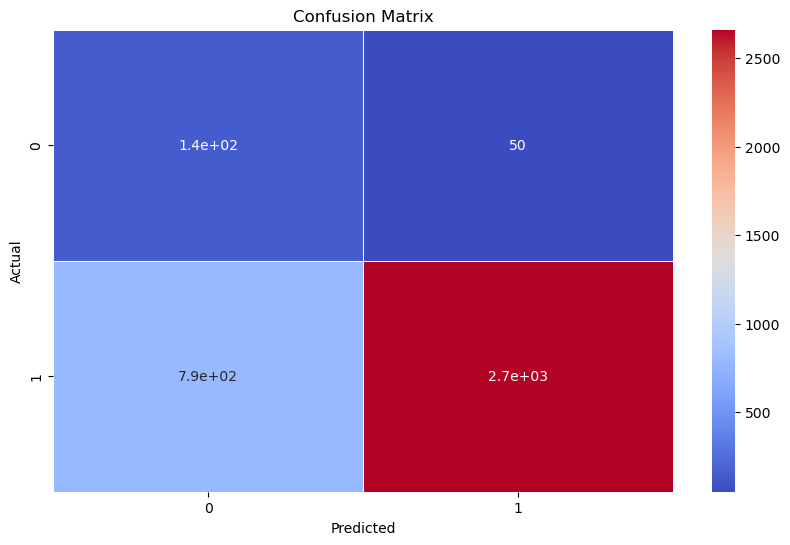
## P1: Linear Regression Imputation + Log Scaling SVM (regularized L1)



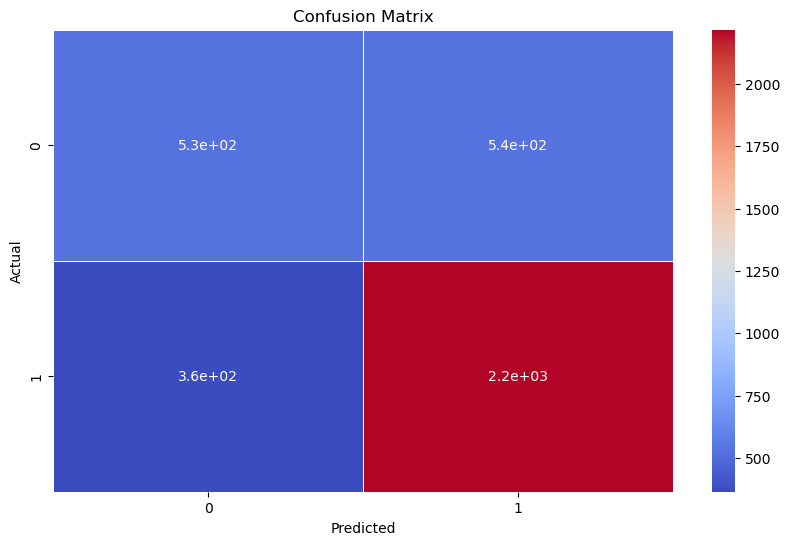
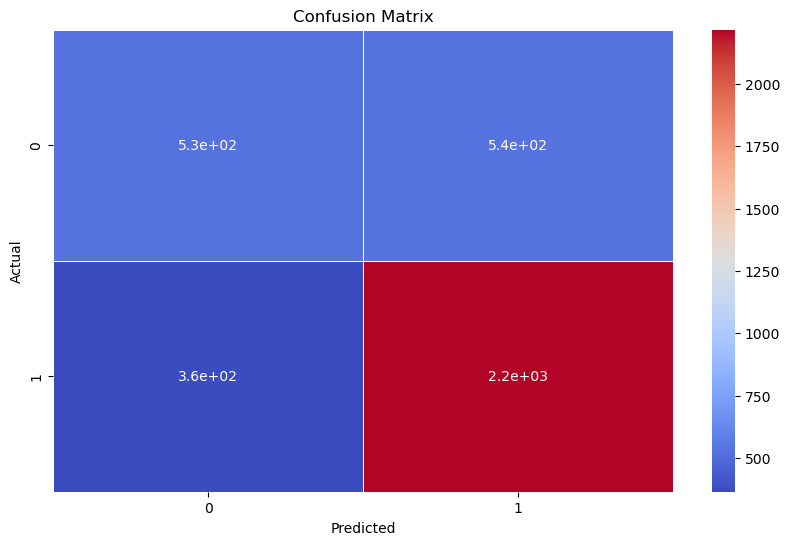
## P2: Linear Regression Imputation + Robust Scaling SVM



## P3: Mean Imputation + Log Scaling SVM (regularized L2)



## P4: Mean Imputation + Robust Scaling SVM



### Discussion on how different preprocessing techniques influenced model performance.

From the previous models as shown above, we see some minor changes on the linear regression and mean imputation on the model but overall the performance of the models almost remain the same in the terms of the loss function, but there were some differences in the confusion matrix for different models. Furthermore, different models took different time execution.

# Final Observations:

## Identify the best-performing model

#### **Logistic:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Accuracy | Precision | Recall | F1-score |
| Pipeline Model P1: Linear Regression Imputation + Log Scaling Logistic Regression (regularized L1) | **69.88** | **26.33** | **10.94** | **15.46** |
| P2: Linear Regression Imputation + Robust Scaling Logistic Regression | **70.63** | **24.68** | **8.54** | **24.68** |
| P3: Mean Imputation + Log Scaling Logistic Regression (regularized L2) | **69.24** | **23.97** | **10.11** | **14.22** |
| P4: Mean Imputation + Robust Scaling Logistic Regression | **69.74** | **28.39** | **9.31** | **14.02** |

##### **P2: Linear Regression Imputation + Robust Scaling Logistic Regression**

##### Reason:

Pipeline 2 for Logistic has good trade-off than all other models present in the pipelines with better performance and execution time.

#### **SVM:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Accuracy | Precision | Recall | F1-score |
| P1: Linear Regression Imputation + Log Scaling SVM (regularized L1) | **78.81** | **64.98** | **34.07** | **44.70** |
| P2: Linear Regression Imputation + Robust Scaling SVM | **76.23** | **52.06** | **64.16** | **57.48** |
| P3: Mean Imputation + Log Scaling SVM (regularized L2) | **76.95** | **74.36** | **15.49** | **25.64** |
| P4: Mean Imputation + Robust Scaling SVM | **75.27** | **49.62** | **59.17** | **53.98** |

**P2: Linear Regression Imputation + Robust Scaling SVM**

##### Reason:

Pipeline 2 for SVM has better performance, accuracy ,recall, precision and F1-score (having a balanced output) than all the models present in the pipeline.

## Explain the impact of imputation and scaling

**Logistic:**

We observe from the tables above, it shows that these models have loss. Using the Linear regression models in P1 and P2 shows the significant accuracy, indicating that imputing these missing values by linear regression preserves feature relationships better than Mean imputation.

Meanwhile, P3 and P4 shows lower accuracy and inconsistent recall, suggesting that replacing missing values may introduce bias and affecting the models performance. Robust scaling in P4 resulted in slightly better precision, making it more effective in handling the outliers/

**SVM:**

We observe from the tables above, it shows P1 achieved the highest accuracy ,meaning log scaling effectively normalized and therefore improved model stability. However, its recall is lower than P2 but has low accuracy. Suggesting that the robust scaling handle the outliers better , improving the recall.

Meanwhile P3 has high precision but low recall and P4 provided the balanced trade-off meaning it stabilize the performance.

## Recommendations on when to use different preprocessing approaches.

**Logistic:**

* Pipeline Model P1: Linear Regression Imputation + Log Scaling Logistic Regression (regularized L1):
  + This model is useful in a scenario where you want to capture as many positive values as possible. As it would be suitable for tasks where missing positive values is more costly than false positives
  + For instance, initial screening process for cancer detection where the goal is to identify as many potential cancer cases as possible.
* P2: Linear Regression Imputation + Robust Scaling Logistic Regression;
  + This model is useful in situations where you need to give priority to the accuracy of the positive values, ignoring some recalls.
  + For example, in the fraud detection for financial transactions, where the priority is to minimize the number of fraudulent transactions that are approved.
* P3: Mean Imputation + Log Scaling Logistic Regression (regularized L2):
  + This model can be useful as a baseline or starting point for further model development and optimization.
  + For instance, in the initial analysis of why customers stop using a service. It helps identify key patterns and factors that contribute to customer churn, making a good initial start.
* P4: Mean Imputation + Robust Scaling Logistic Regression:
  + This model can be useful in those situations where you want to maintain a reasonable accuracy across both positive and negative values.
  + For instance, when we want to predict either an employee might leave a company or not. By making a good balance in the model, it can be helpful for the HR.

**SVM:**

* P1: Linear Regression Imputation + Log Scaling SVM (regularized L1)
  + This model is useful in a scenario where you want to capture as many positive values as possible. Therefore, accepting the higher number of false positives.
  + For instance, initial screening process for cancer detection where the goal is to identify as many potential cancer cases as possible.
* P2: Linear Regression Imputation + Robust Scaling SVM:
  + This model is useful in situations where you need to give priority to the accuracy of the positive values, ignoring some recalls.
  + For example, in the fraud detection for financial transactions, where the priority is to minimize the number of fraudulent transactions that are approved.
* P3: Mean Imputation + Log Scaling SVM (regularized L2):
  + This model can be useful as a baseline or starting point for further model development and optimization.
  + For instance, in the initial analysis of why customers stop using a service. It helps identify key patterns and factors that contribute to customer churn, making a good initial start
* P4: Mean Imputation + Robust Scaling SVM:
  + This model can be useful in those situations where you want to maintain a reasonable accuracy across both positive and negative values.
  + For instance, when we want to predict either an employee might leave a company or not. By making a good balance in the model, it can be helpful for the HR.