

Census Income

BY

GOPALAKRISHNAN KUMAR

Objectives

To perform a predictive task of classification to predict whether an individual makes over 50K a year or less using machine learning algorithms

Data Preparation

- CENSUS DATASET WAS PROVIDED WITH VARIABLES SUCH AS AGE, WORK-CLASS, EDUCATION, NO OF YEARS OF EDUCATION, MARITAL STATUS, OCCUPATION AND RELATIONSHIP.
 - LAB ENVIRONMENT- R STUDIO
 - DOMAIN SOCIAL
-

Data Preprocessing

All the missing values are replaced with NA.

All the rows that contains NA values are removed

All whitespaces from the columns are removed

Data Manipulation

To analyze the dataset, various functions from the **dplyr** package

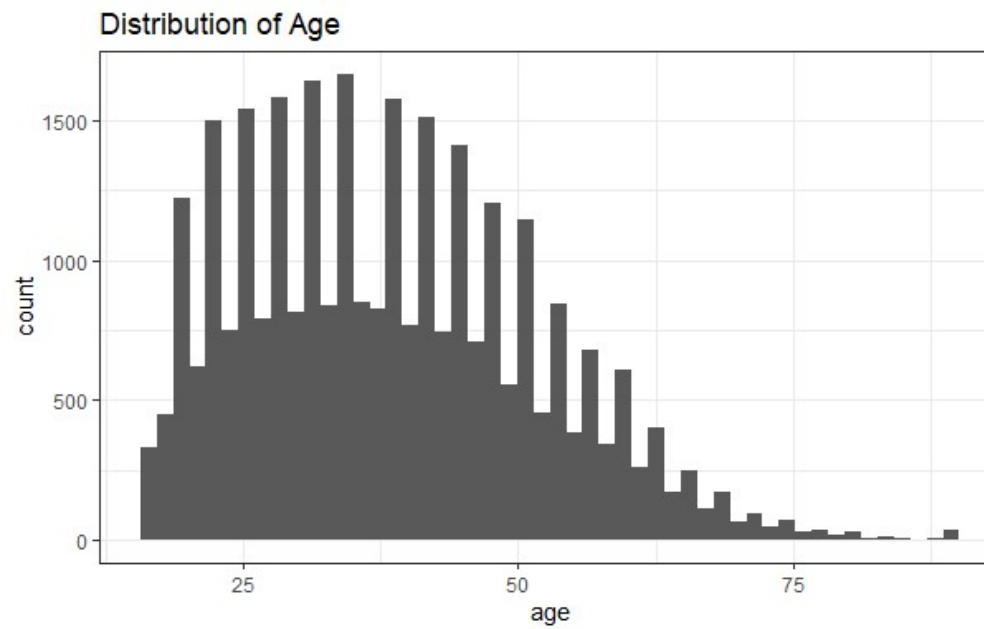
Data Visualization

Helps to understand patterns, trends and outliers in dataset.

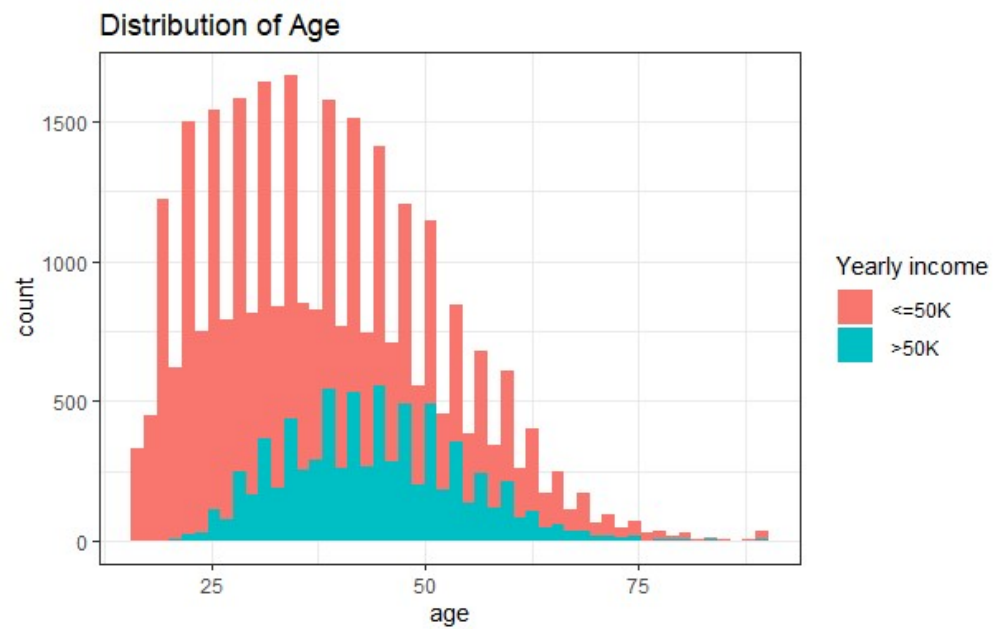
In this project, data visualization has been performed in the form of

- ❖ Bar graph
- ❖ Histogram
- ❖ Scatterplot
- ❖ Boxplot

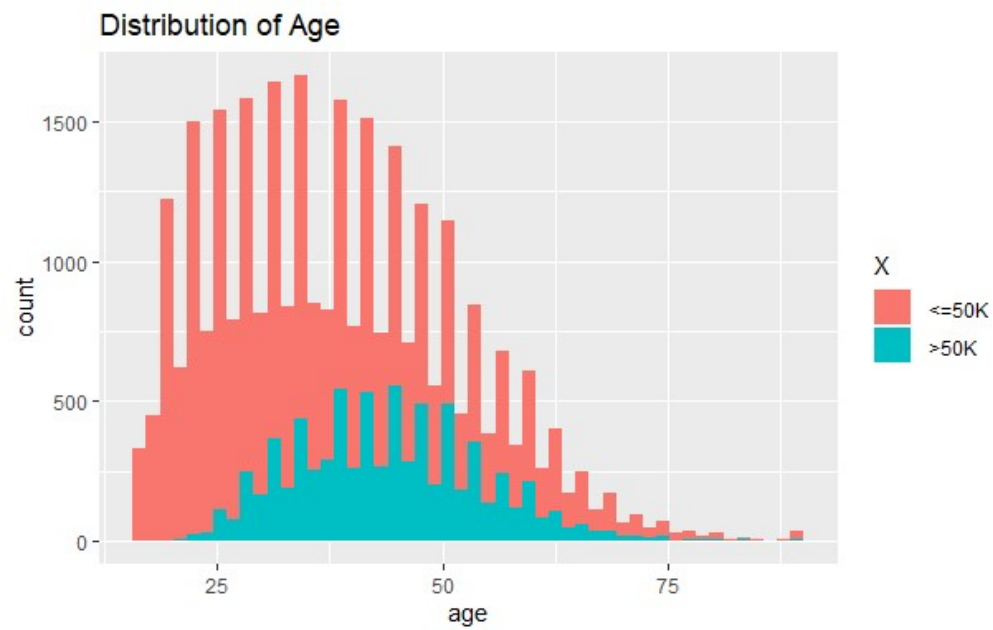
Bar Graph



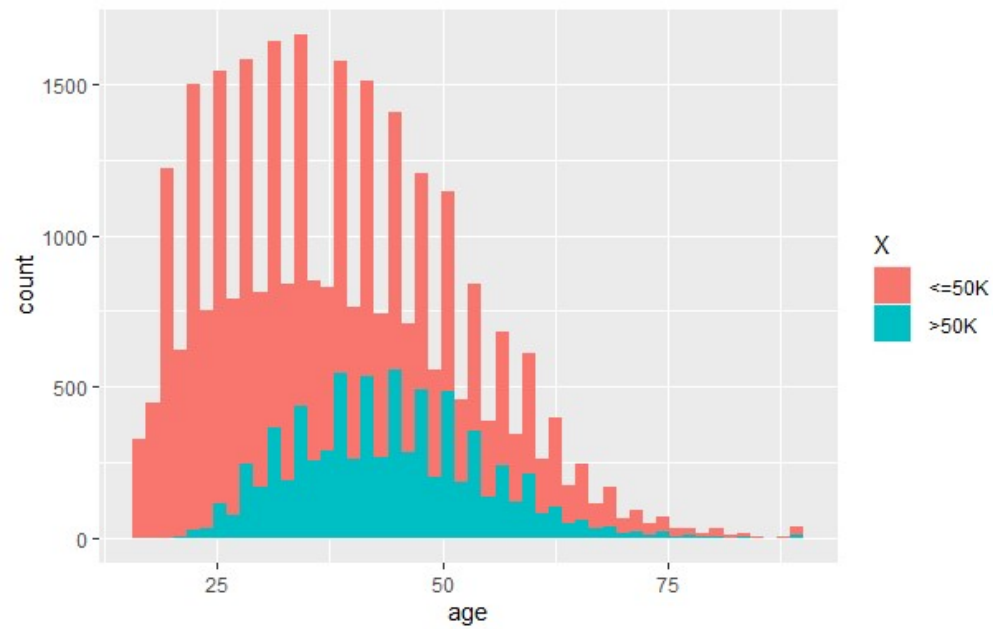
Bar Graph



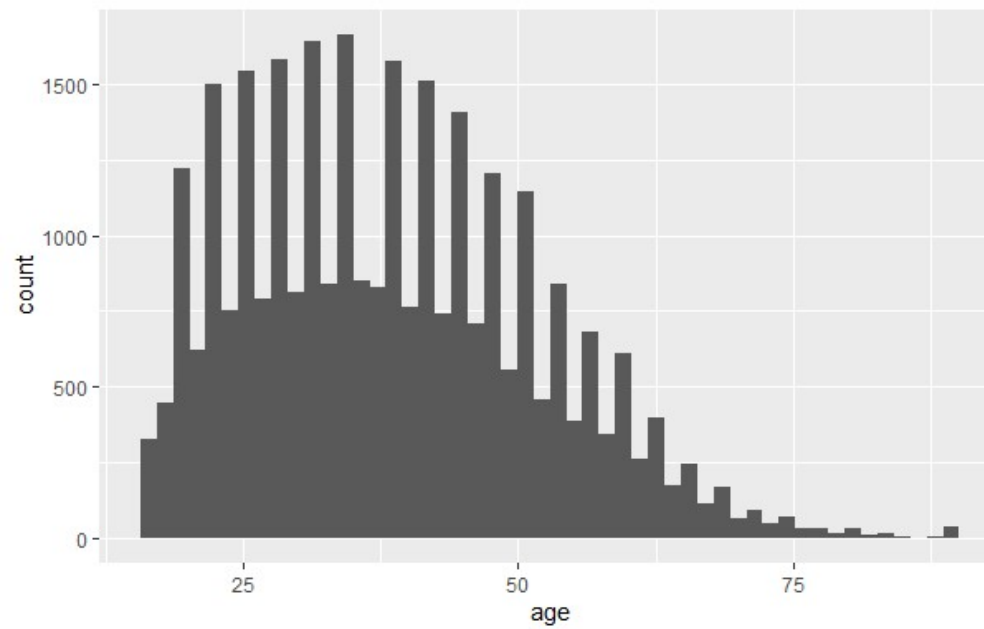
Bar Graph



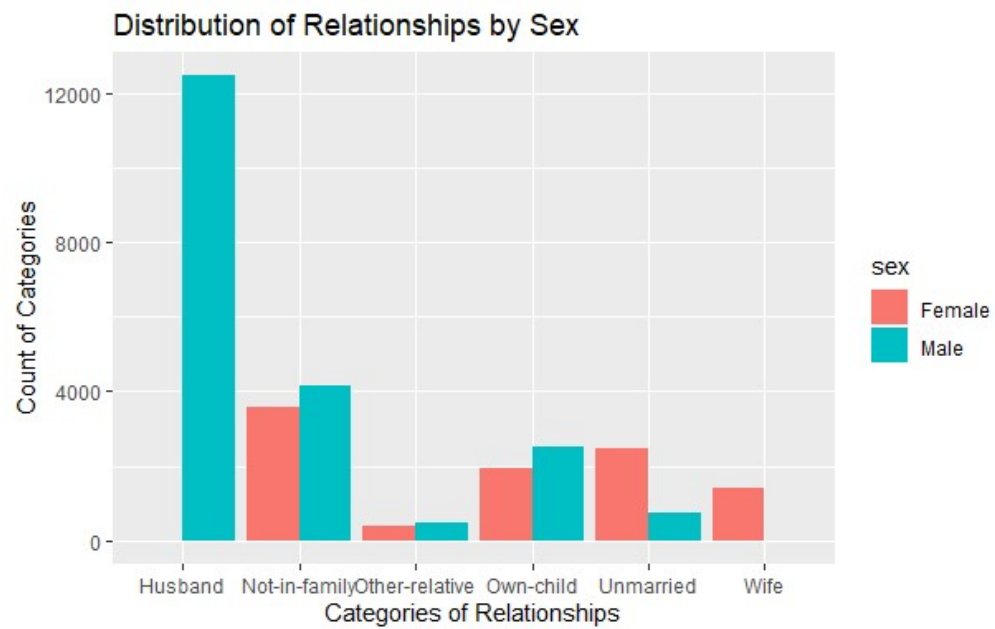
Bar Graph



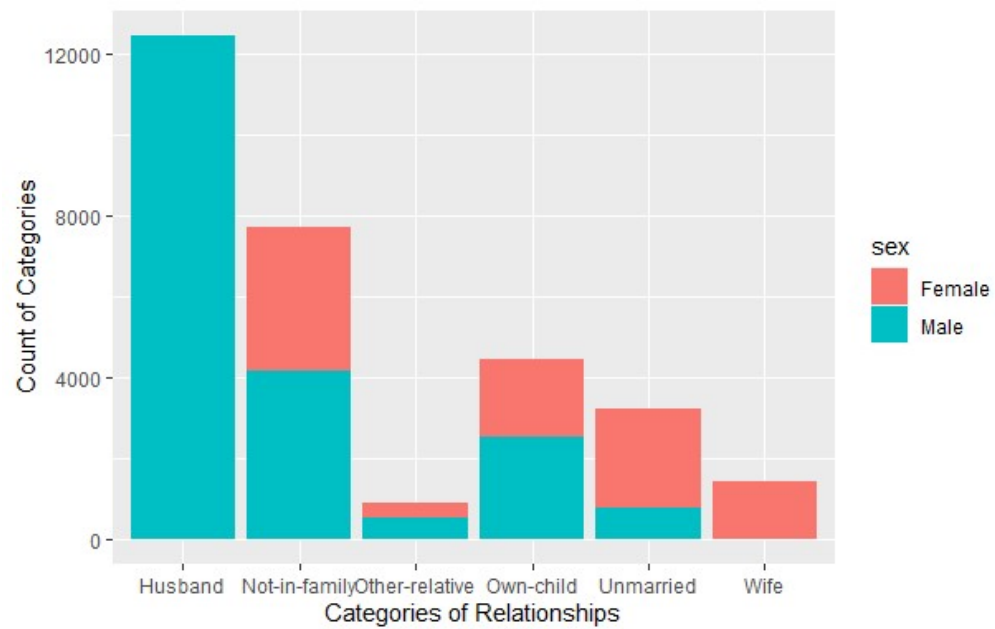
Bar Graph



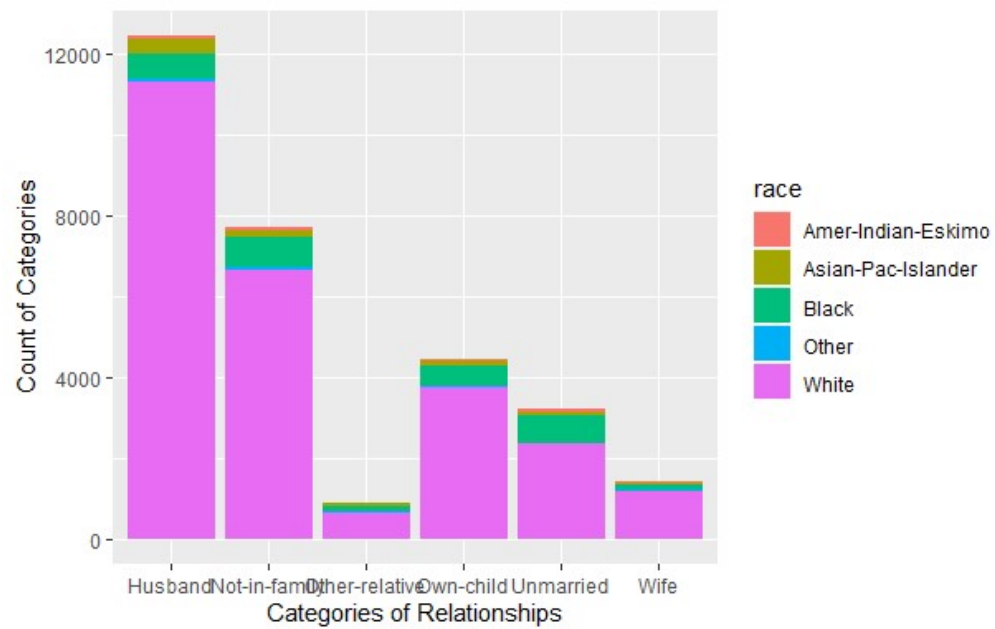
Histogram



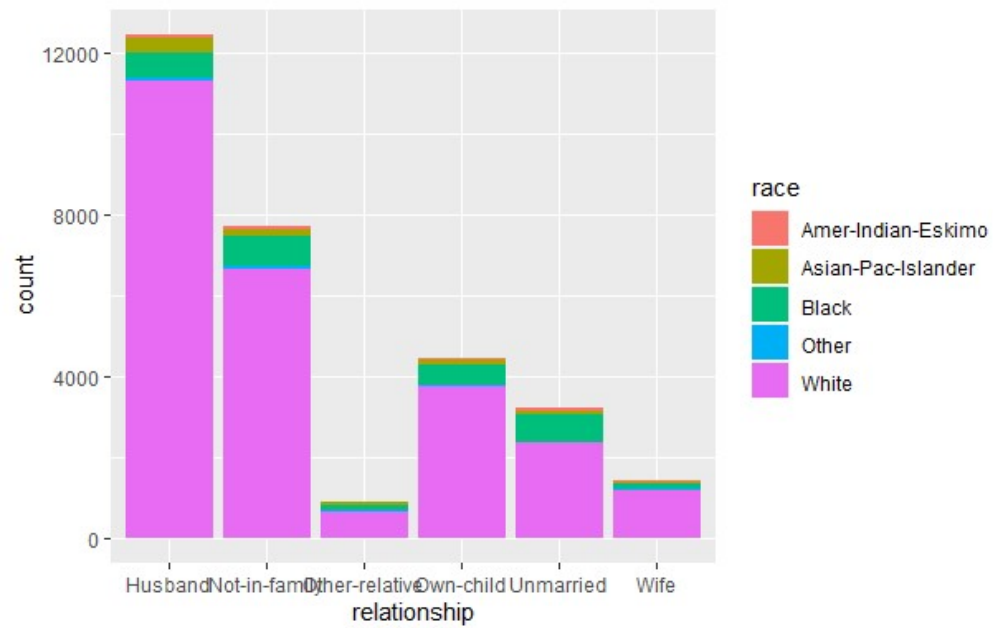
Histogram



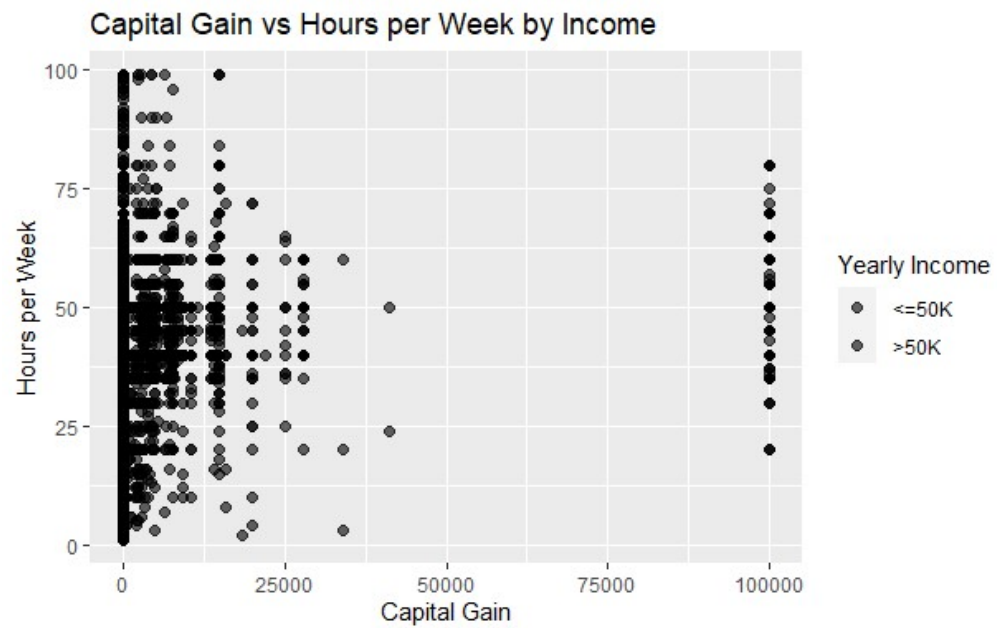
Histogram



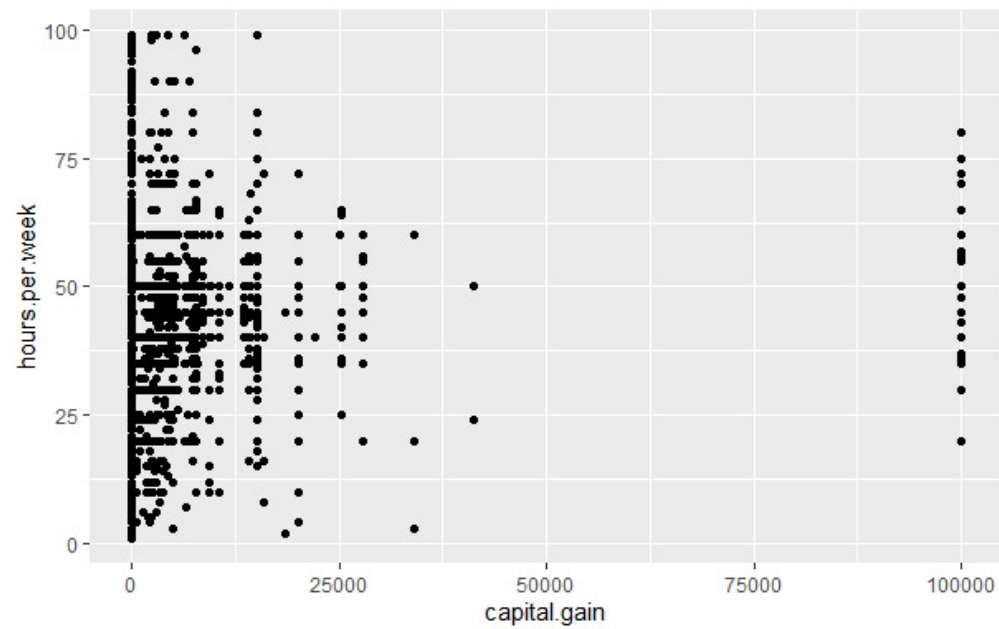
Histogram



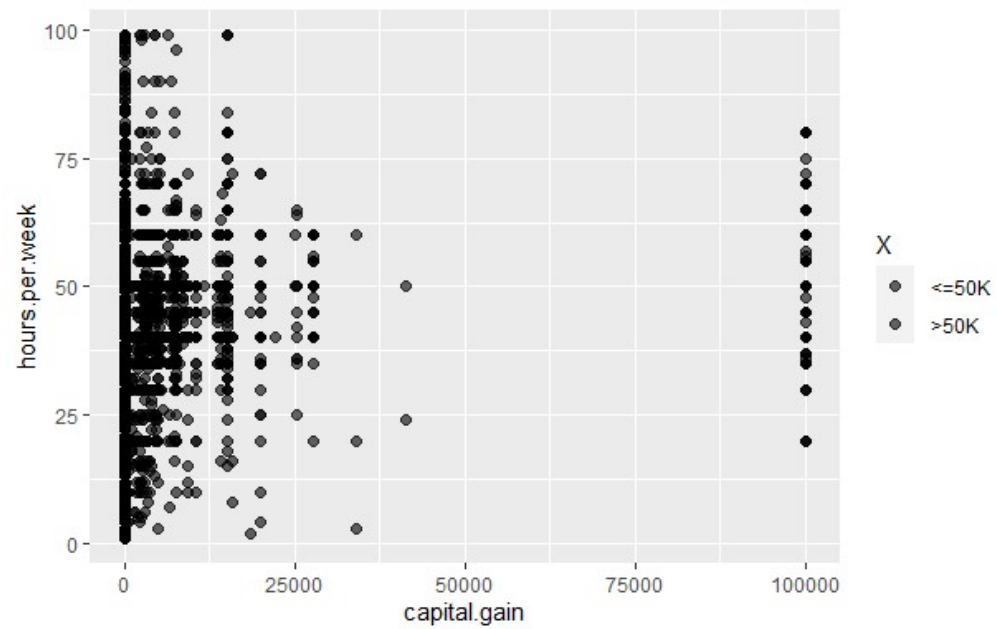
Scatterplot



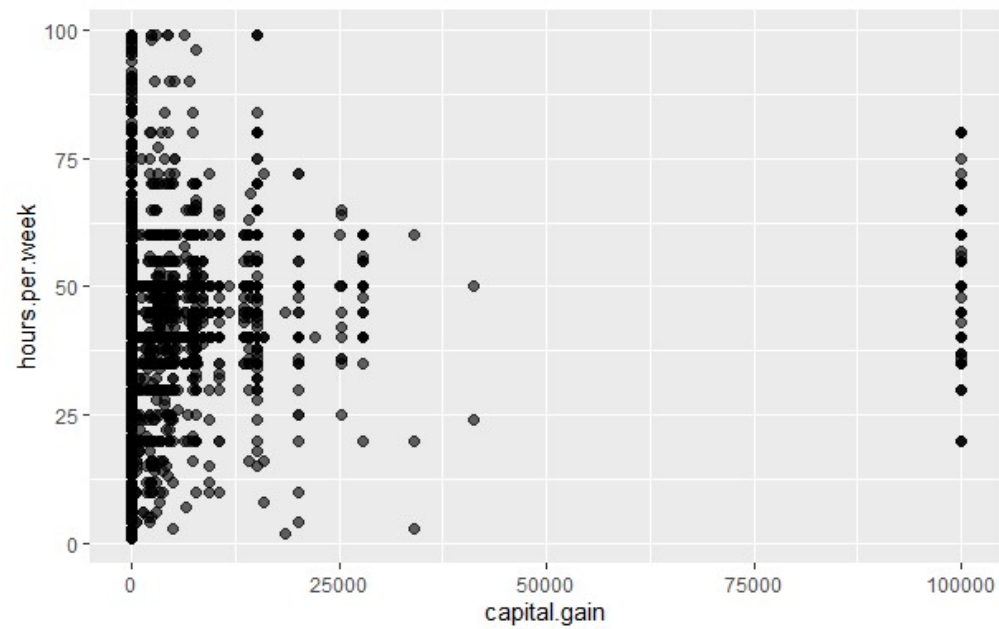
Scatterplot



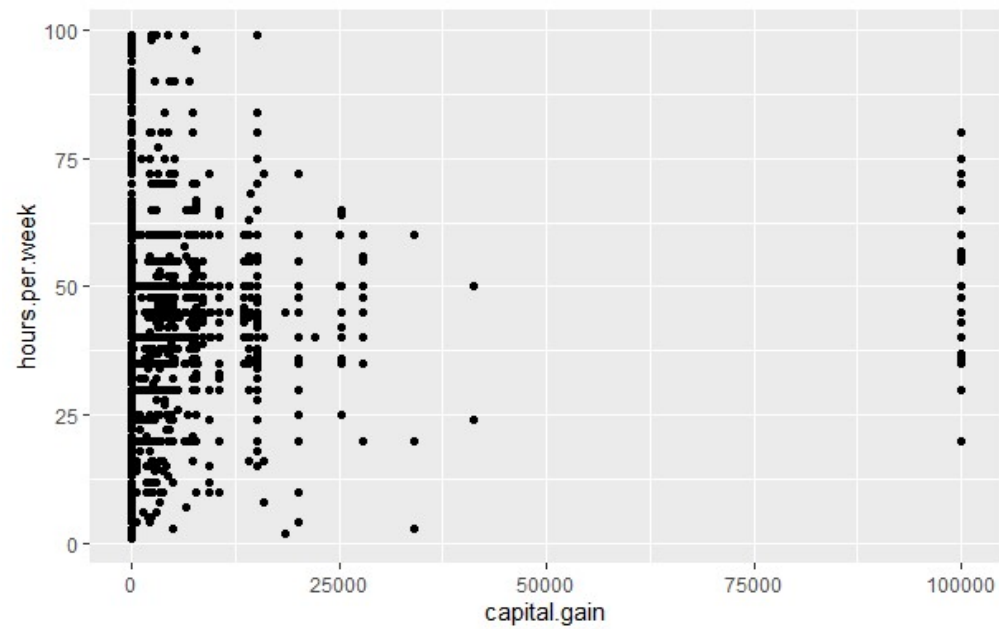
Scatterplot



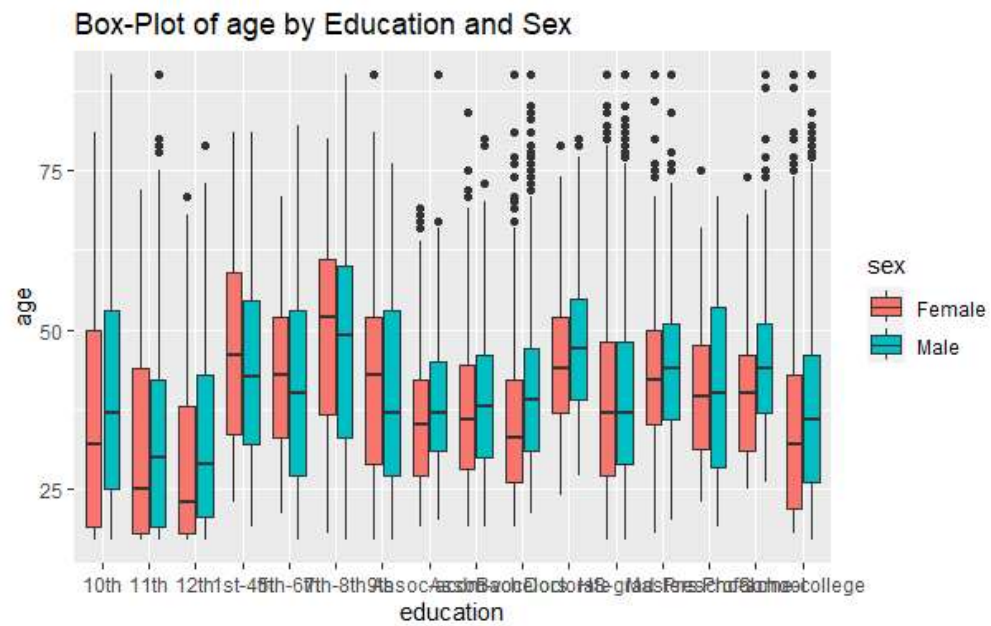
Scatterplot



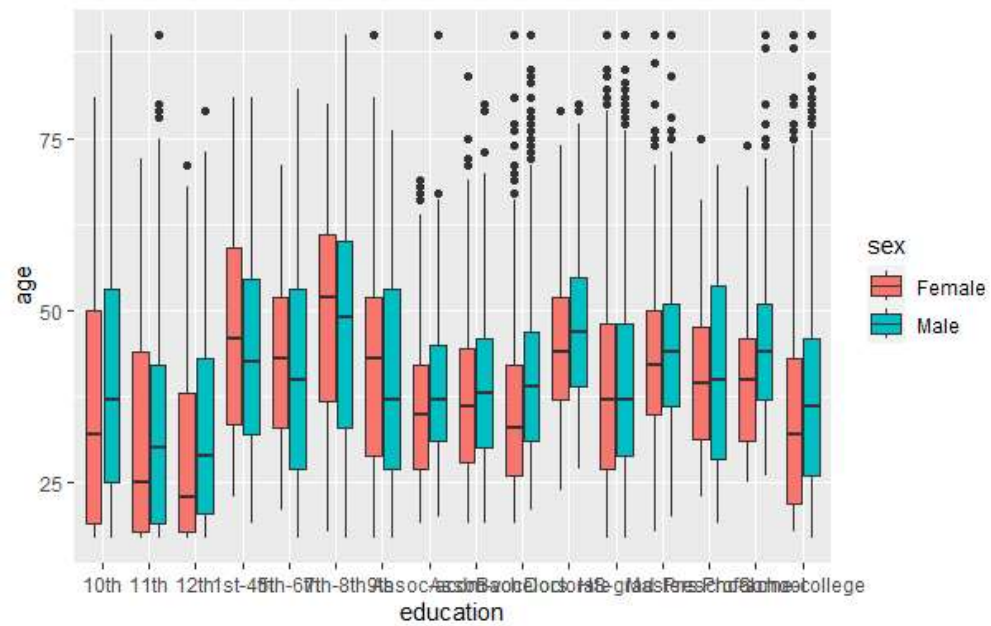
Scatterplot



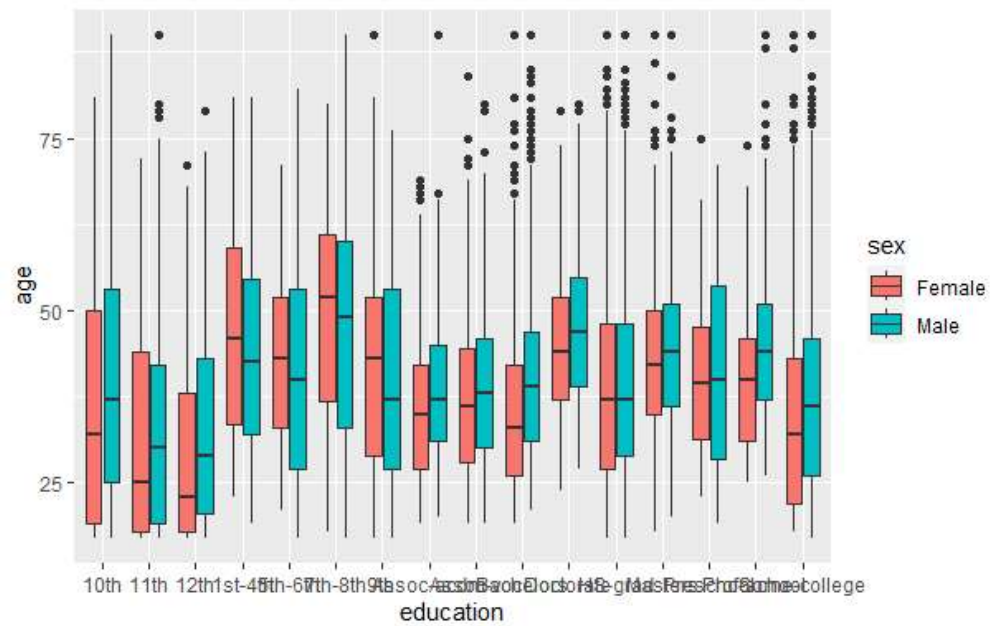
Boxplot



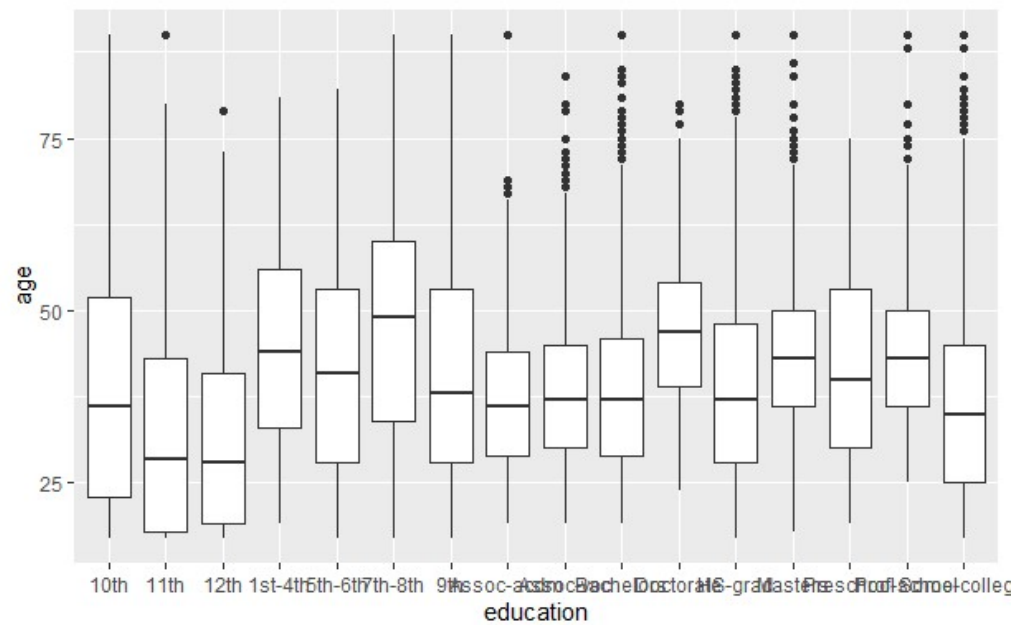
Boxplot



Boxplot



Boxplot



Model Building

Decision Tree model has been applied to predictive task of classification to predict whether an individual makes over 50K a year or less

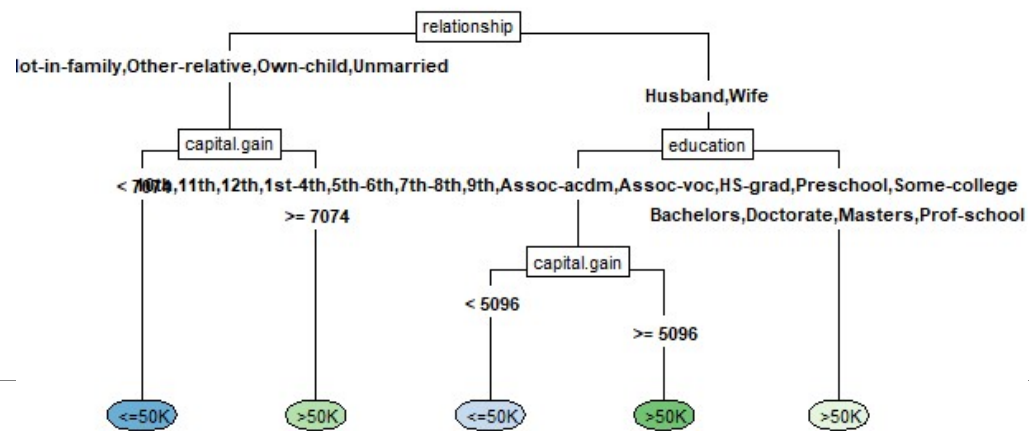
Dataset has been divided into training and testing sets in 70:30

A Decision Tree model was built with dependent variable as yearly income (X-axis) and the rest of variables as independent variables

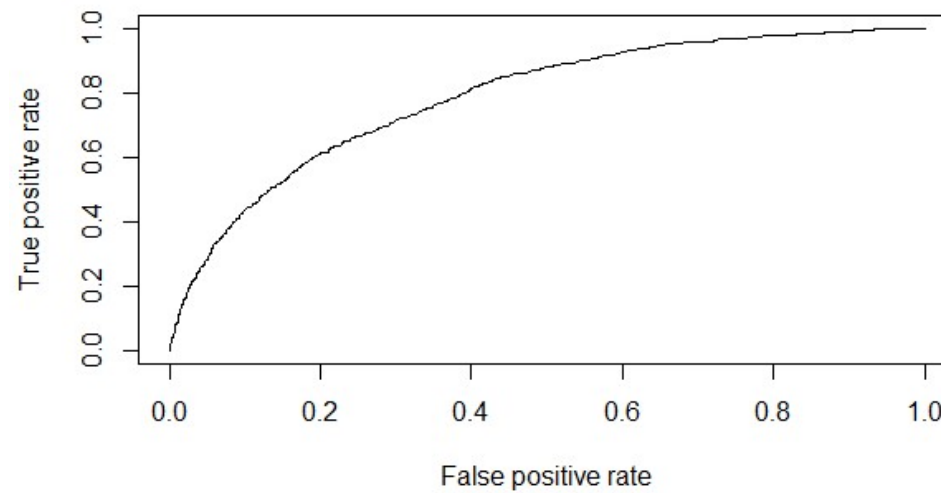
The values are predicted on the test set

Confusion has been built to calculate the accuracy

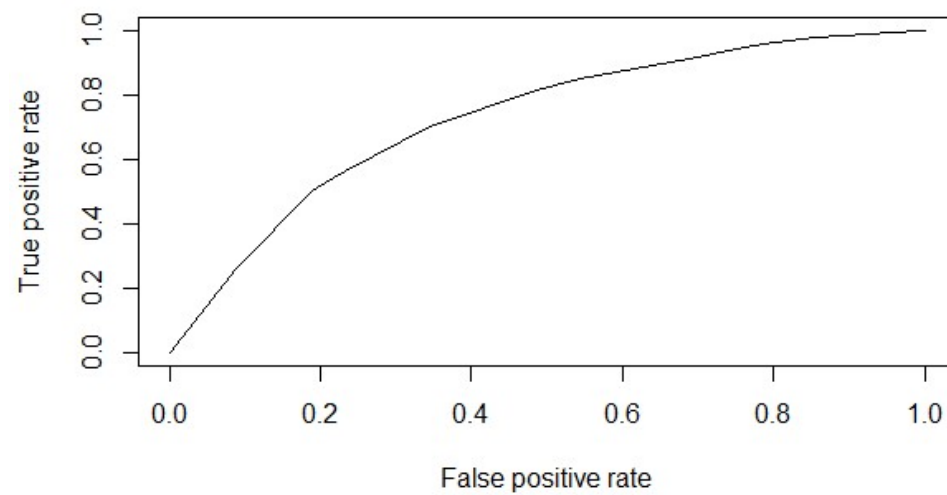
Decision Tree



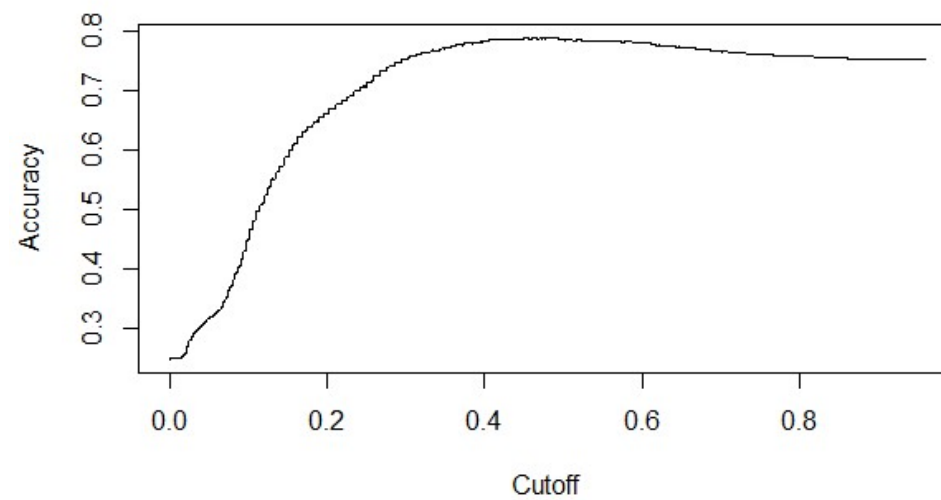
Confusion Matrix



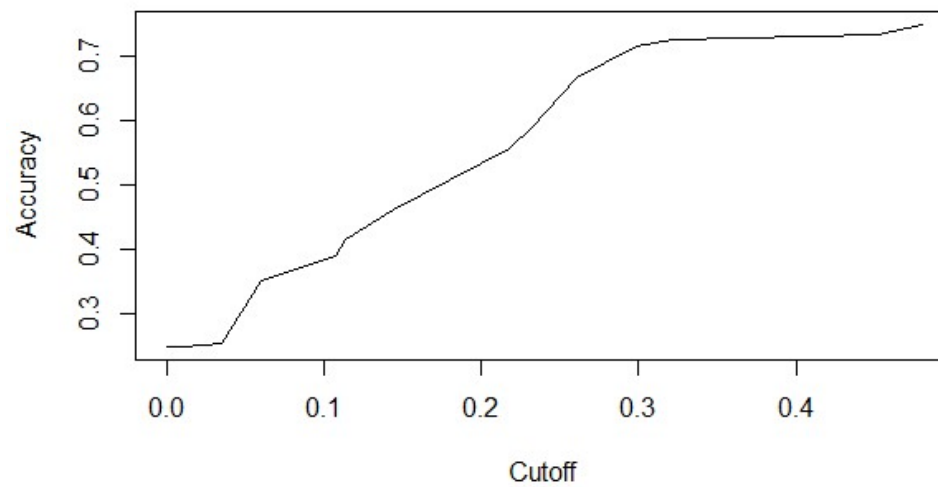
Confusion Matrix



Confusion Matrix



Confusion Matrix



Confusion Matrix

