TASK – 1

PROBLEM STATEMENT:

Explore the given data set with EDA techniques and build a suitable model for predicting whether the salary of the person is >50k or not and visualize the results. The algorithm used for the model should be built from scratch.

OBJECTIVE:

- Interpretation
- Exploratory Data Analysis
- Visualization
- Data Pre-processing
- Model Building
- Validation

INTERPRETATION:

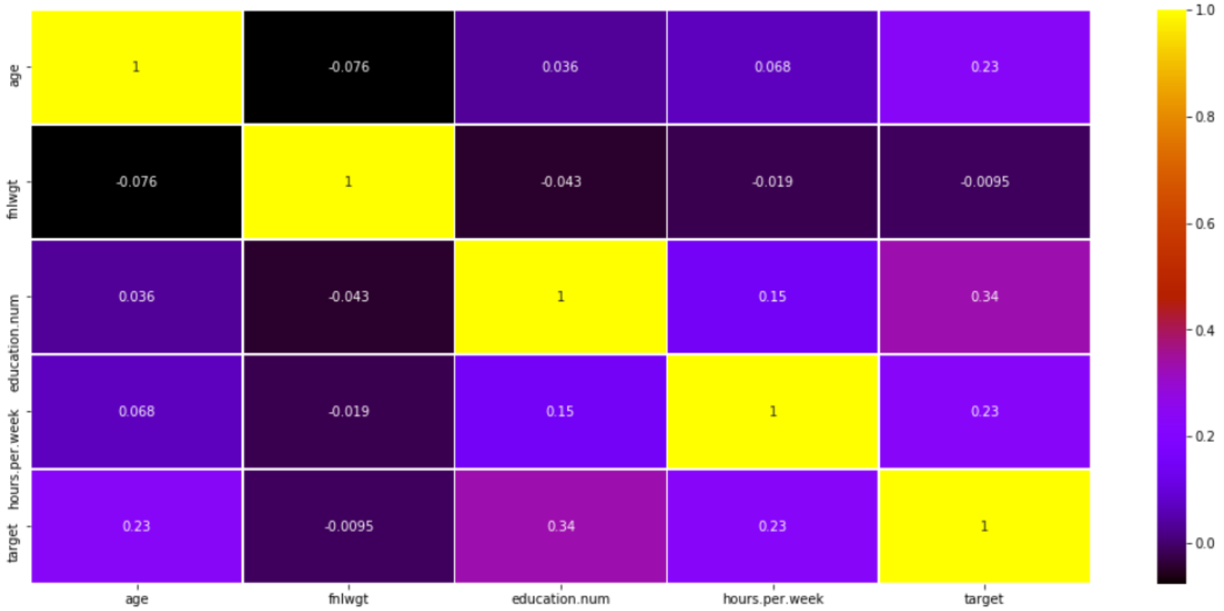
- The dataset contains train and test file.
- The train dataset contains 32561 rows and 15 columns
- The test dataset contains 16281 rows and 15 columns
- The target attribute is of binary classification

EXPLORATORY DATA ANALYSIS:

- Handling of missing values:
    - Most of the columns which has missing values are categorical
    - So, the missing values are handled by imputation of mode.
- Removal of zero-valued columns:
    - The attributes named [‘capital.gain’, ‘capital.loss’] contains 75% of zero(s) and these two can be removed.
- Removal of duplicate values:
    - There are 26 rows duplicated in the dataset. It is removed using drop_duplicates.
- Target Attribute:
    - The target attribute contains two values ‘<=50K’, ‘>50K’
    - The string values are converted into binary format using Label Encoding

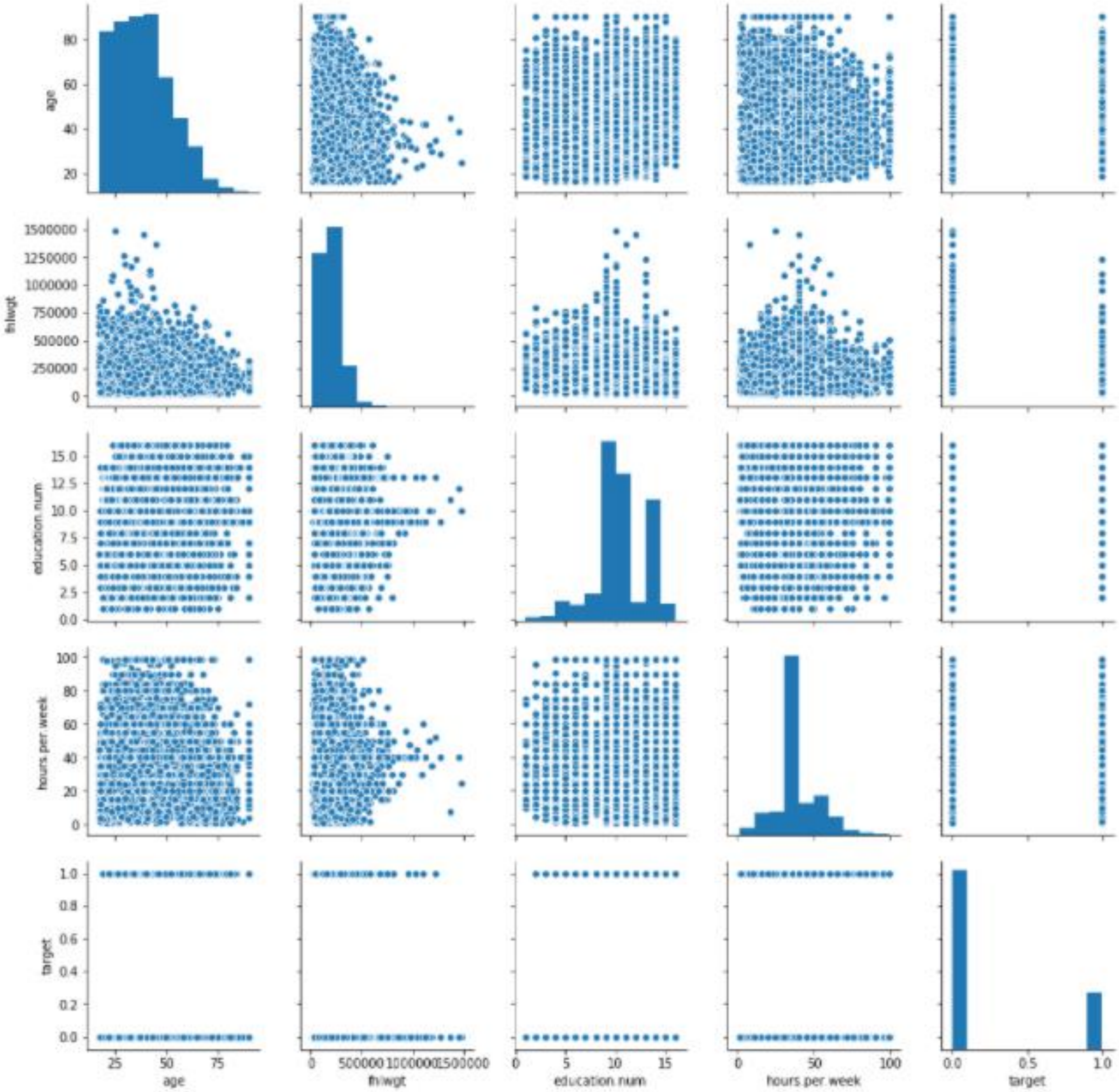
- Feature Attributes:
  - The feature columns of the dataset are categorical data. It is transformed into numeric values using label encoding function.

VISUALIZATION:

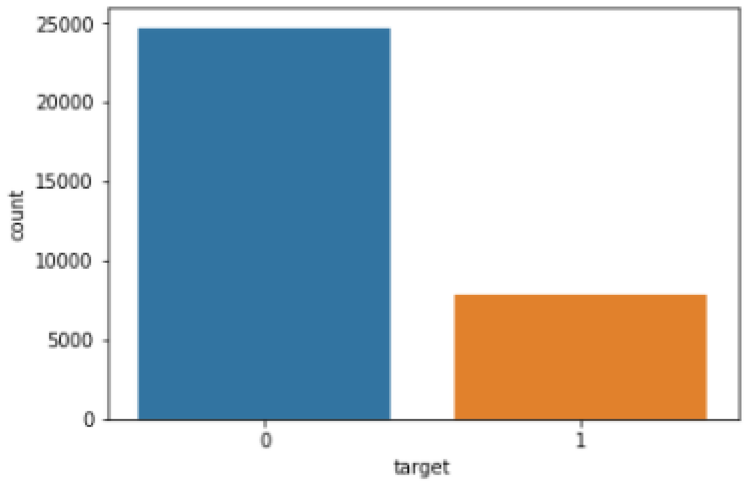


Heatmap also known as Correlation map explains the correlation between the features

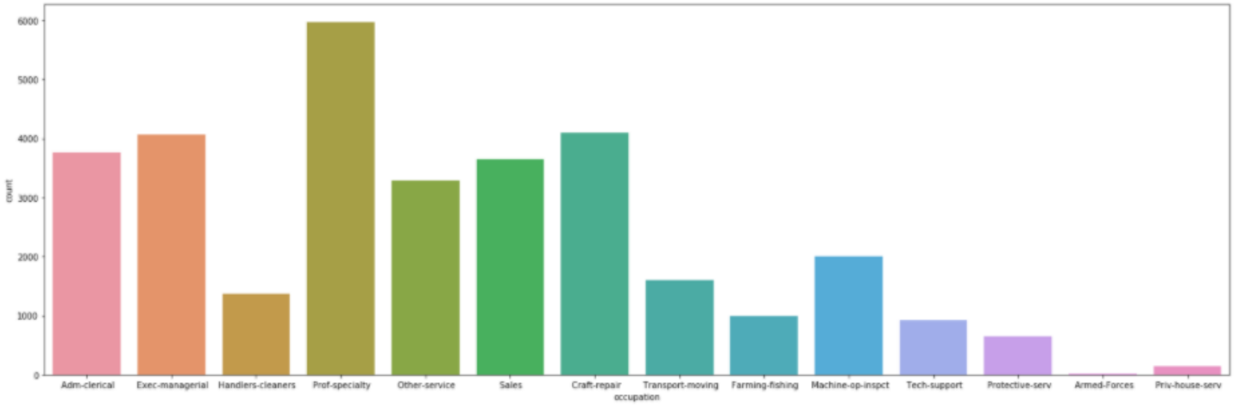
```
<Figure size 720x360 with 0 Axes>
```



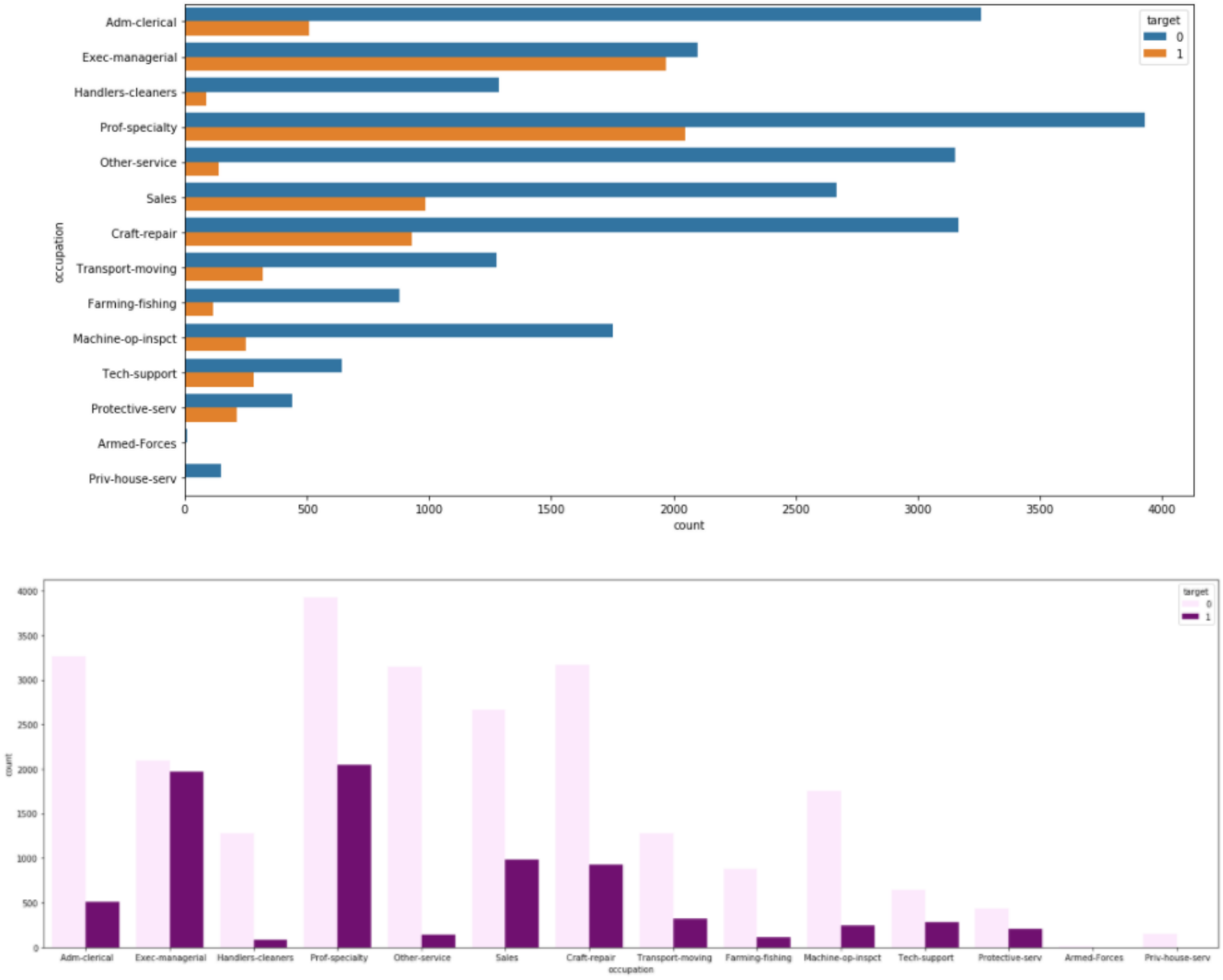
Pair plot of the provided training data



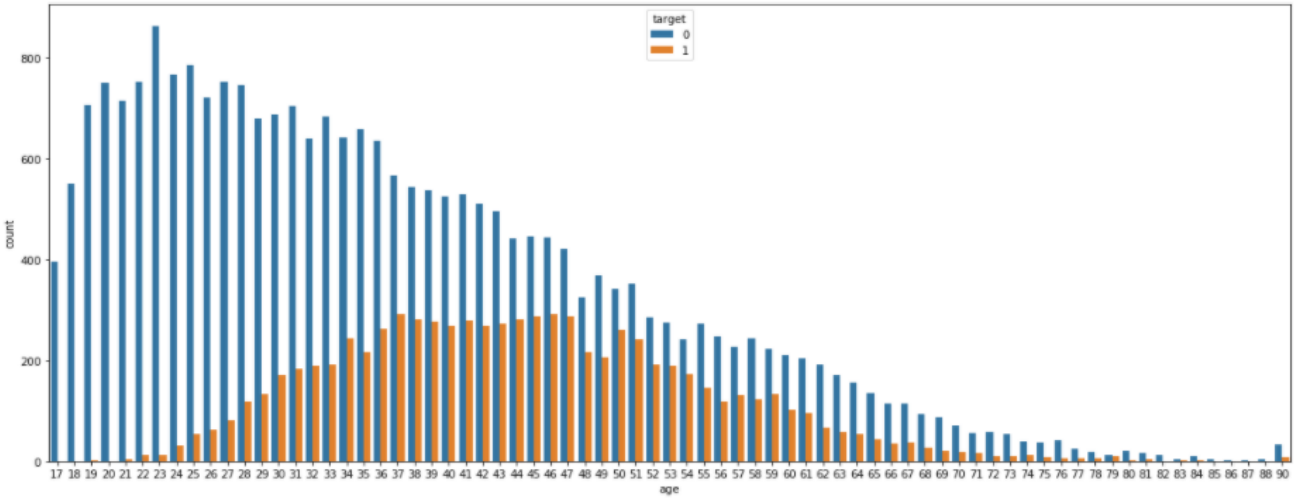
Count of people whose salary is ‘<=50K’ labelled as 0, ‘>50’ labelled as 1



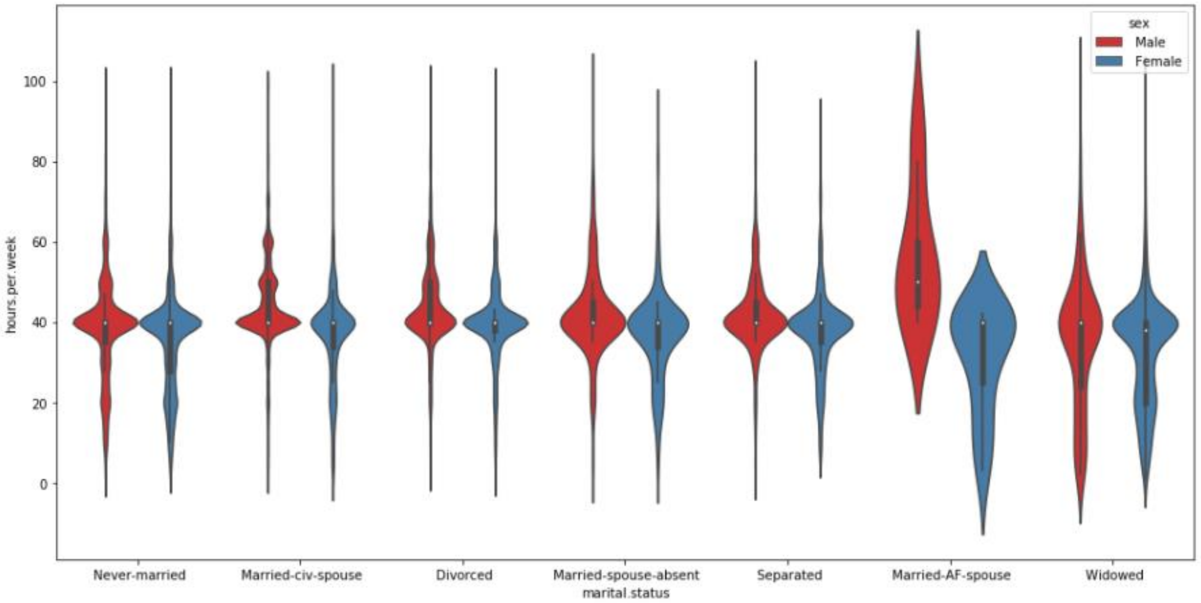
The count plot of the people working at different sectors



The count plot of the occupations with respect to the target attribute



The count of different ages of people with respect to the target attribute



Violin plot which explains the data distribution of marital status and hours per week worked with respect to the target attribute

DATA PRE-PROCESSING:

The dataset is scaled using MinMax Scaler and the values are passed on to the model

MODEL BUIDLING:

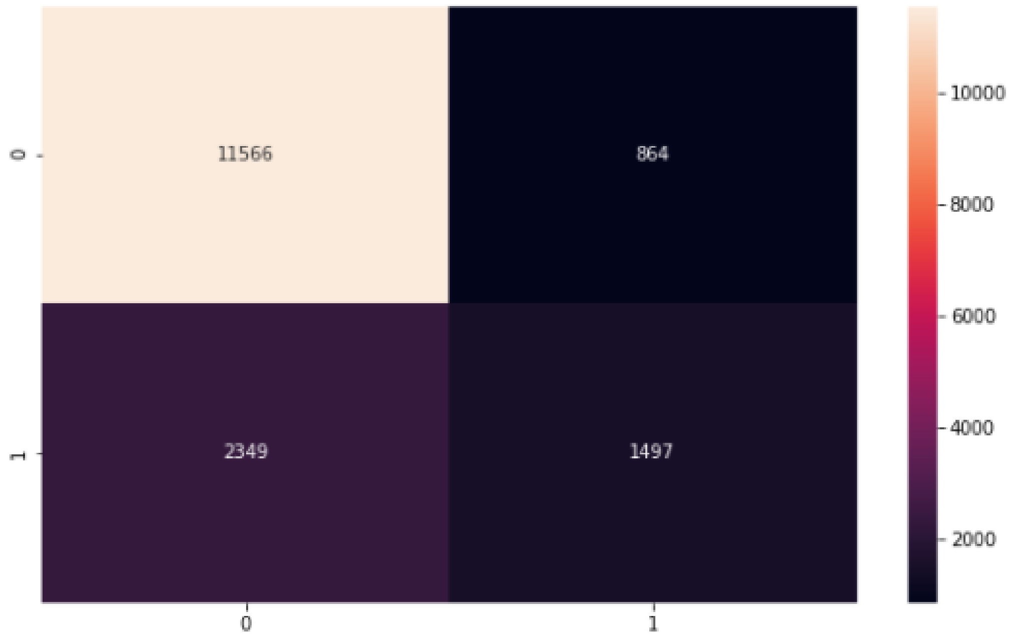
For the provided dataset, Logistic Regression model is fitted. Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.

The model is built from scratch without the use of any libraries. The test values are also predicted using the model.

VALIDATION:

The model acquired an accuracy of 80.26%

The confusion matrix of the model is provided below:



The model built from scratch and the model built using sklearn acquired a similar accuracy of 80.26%.