Analysis on earning potential on Adult data set

Author

Krishna Sai Patha(s3670773)

Avinash Mada (s3674854)

Abstract

The aim of this report was to identify individuals whose salary exceeds a specified value based on demographic information such as age, level of education, employment type and working hours per week. The process involved in exploring the data set, preparing and modelling the data set was discussed. Topics such as the statistical analysis on the attributes, outliers and missing values detection, data transformation and performance analysis of the proposed classifiers were discussed. Overall, the results indicate that most of the individuals are earning less than $50,000 per year. The report concludes that most of the individuals are earning less than $50,000 per year.

Table of Contents

1.Introduction

This assignment will investigate on the demographic attributes of the adult income data set in order to create one or more classification models which are capable in identifying whether an individual can earn more or less than US $50,000. This data set was collected from the United States Census Bureau database and referred to as Adult data set. This data set contains 13961 rows with following attributes.

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. income: >50K, <=50K (indicating an individual can earless or more than US$50,000)

This report will describe the work carried out during data preparation, modelling and evaluation including data formatting or other quality issues and implementation of various classifiers to evaluate and achieve required results.

2. Method

The study was conducted by removing the outliers in the data and fixing all the missing values and creating many classification models which can predict individuals whose salary exceeds fifty thousand US dollars by analysing census data containing demographic information such as age, gender, education level and employment type. The income attribute in the census data is the target attribute in our analysis. The analysis gives a clear description about various statistics of the attributes.

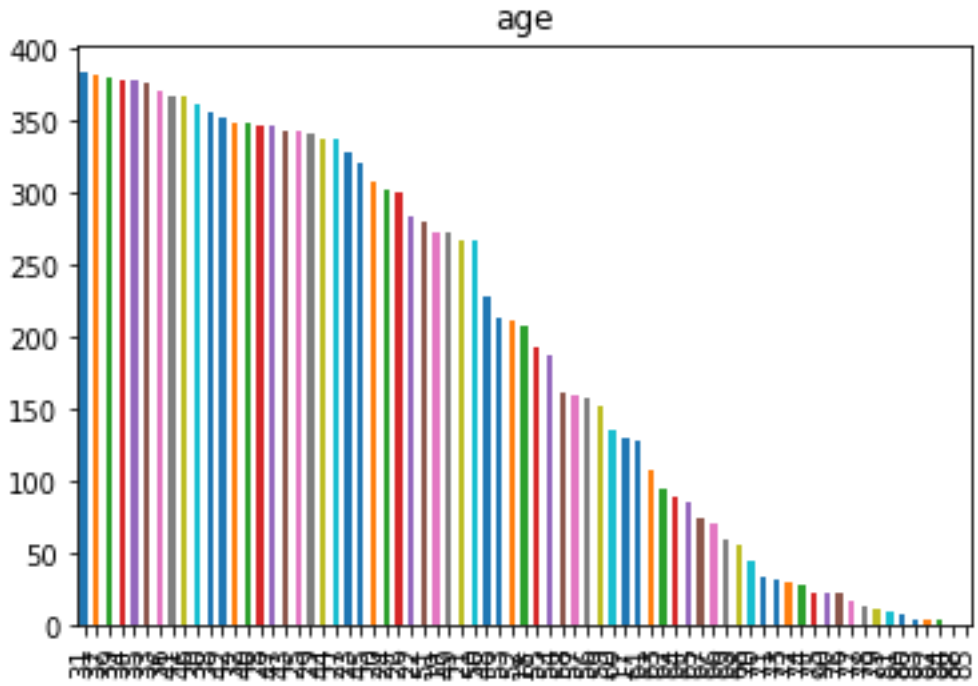
3. Results

The adult dataset is loaded into pandas with appropriate libraries and removed all missing values and outliers. Here is a small visual of the data set.

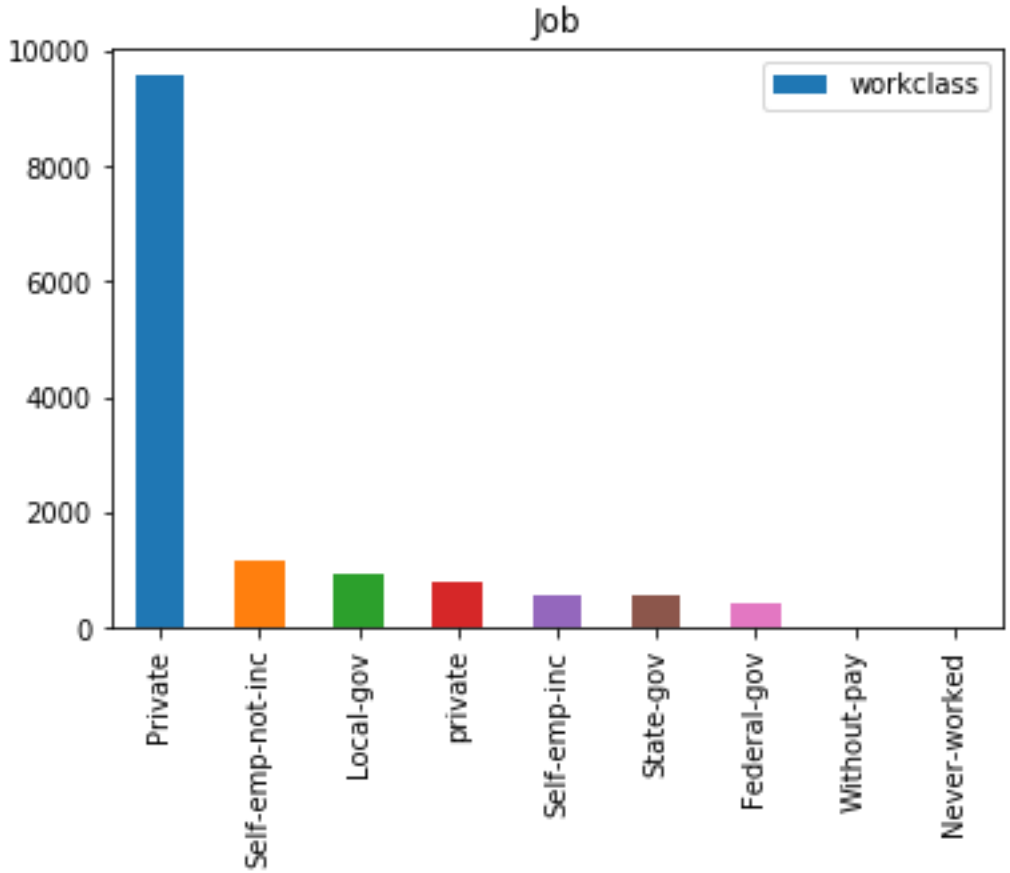
	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	na
7474	46	Self-emp-not-inc	236852	Bachelors	13.0	Married-civ-spouse	Sales	Husband	White	Male	0.0	0.0	45.0	
339	31	Private	184306	Assoc-voc	11.0	Never-married	Transport-moving	Own-child	White	Male	0.0	1980.0	60.0	
11569	41	Private	213055	Assoc-acdm	12.0	Never-married	Adm-clerical	Not-in-family	Other	Female	0.0	0.0	50.0	

Here are the visualizations of all the attributes in the dataset.

Age :

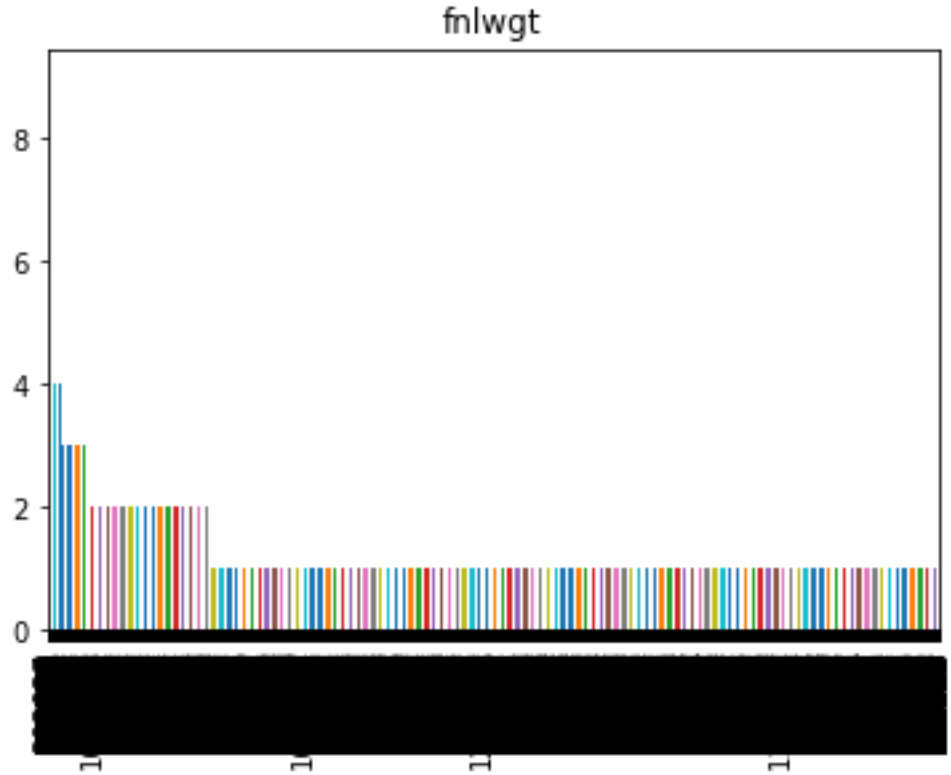


Work Class:



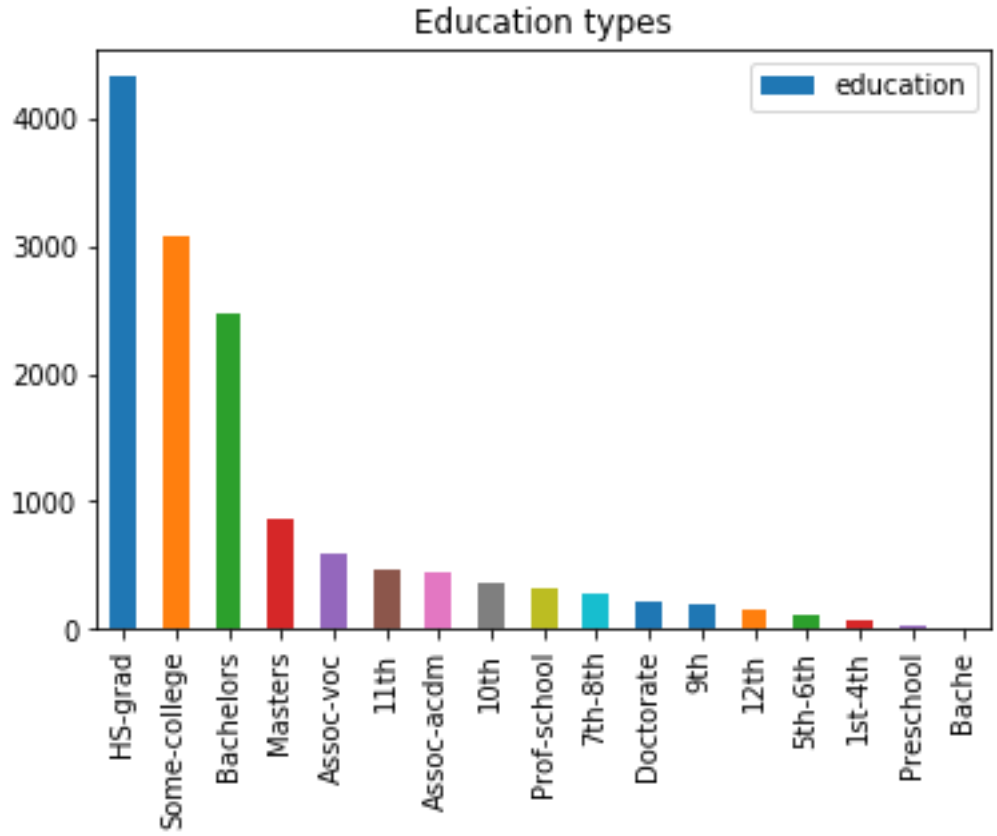
From the figure we can say that there are more number of private employs and self employee takes second position.

Fnlwgt:

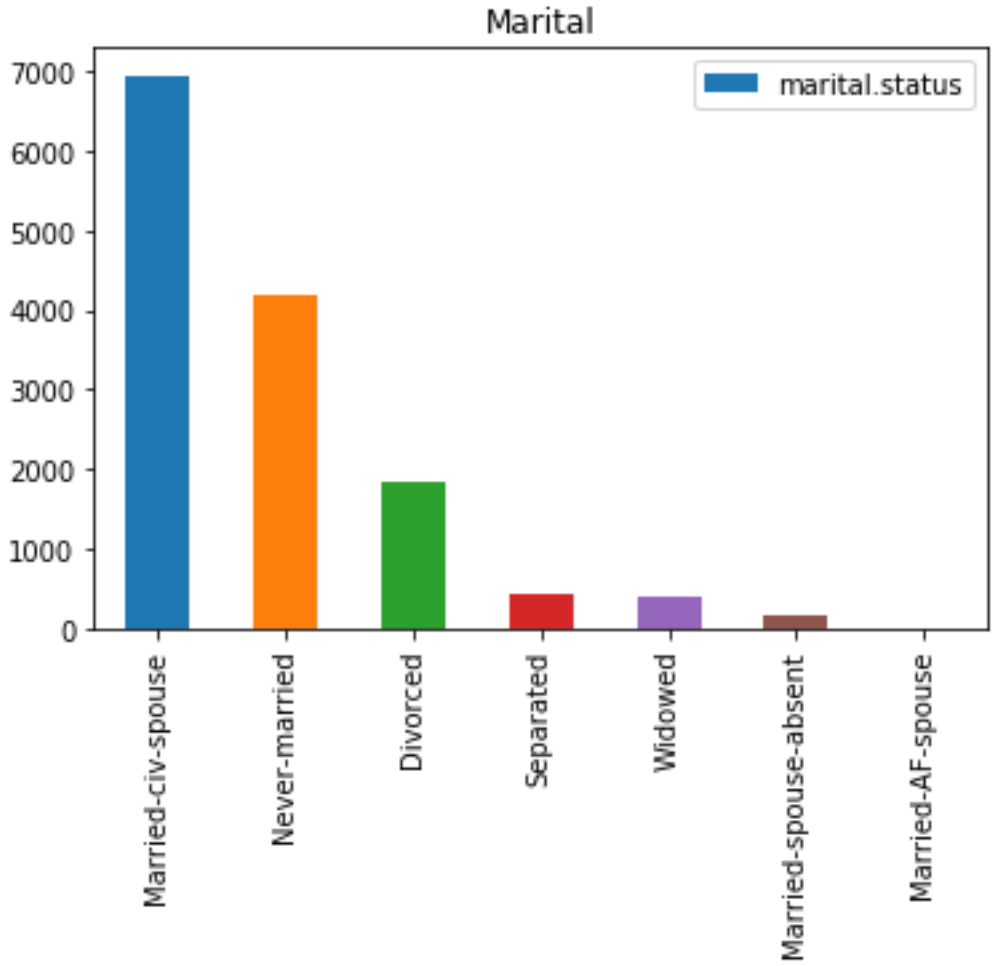


Fnlwgt is final weights on Current Population Surveys

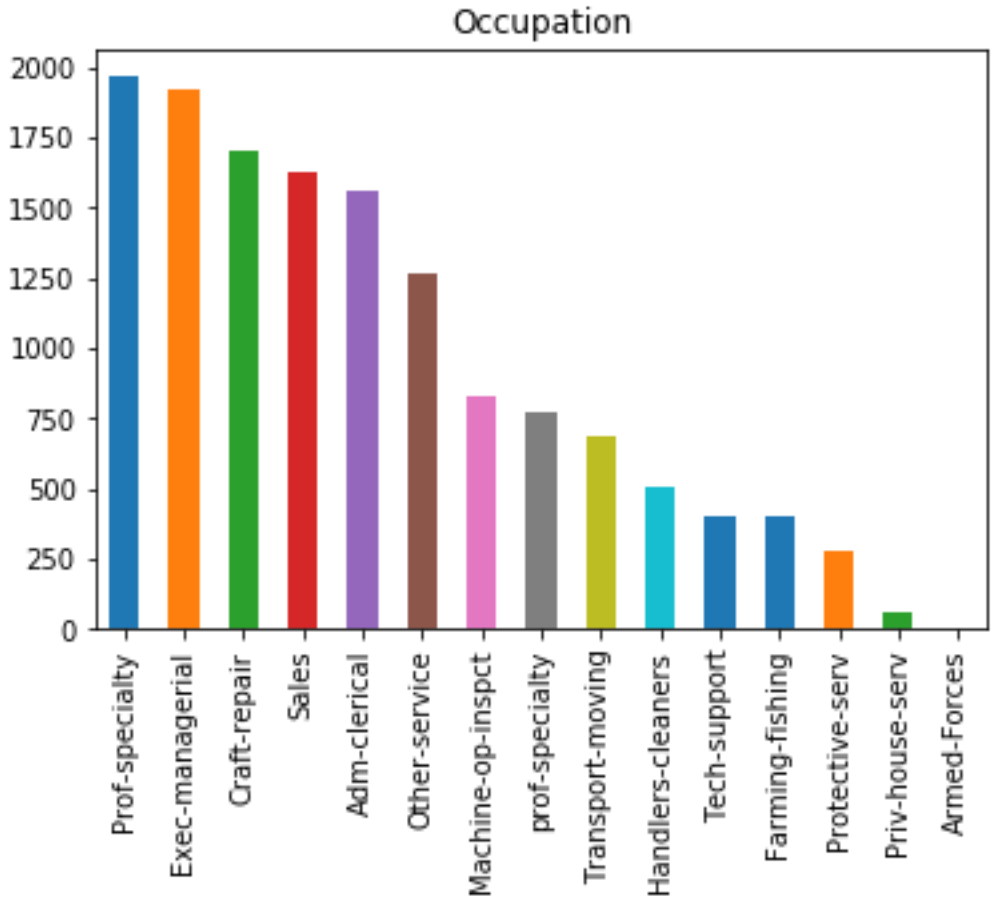
Education:



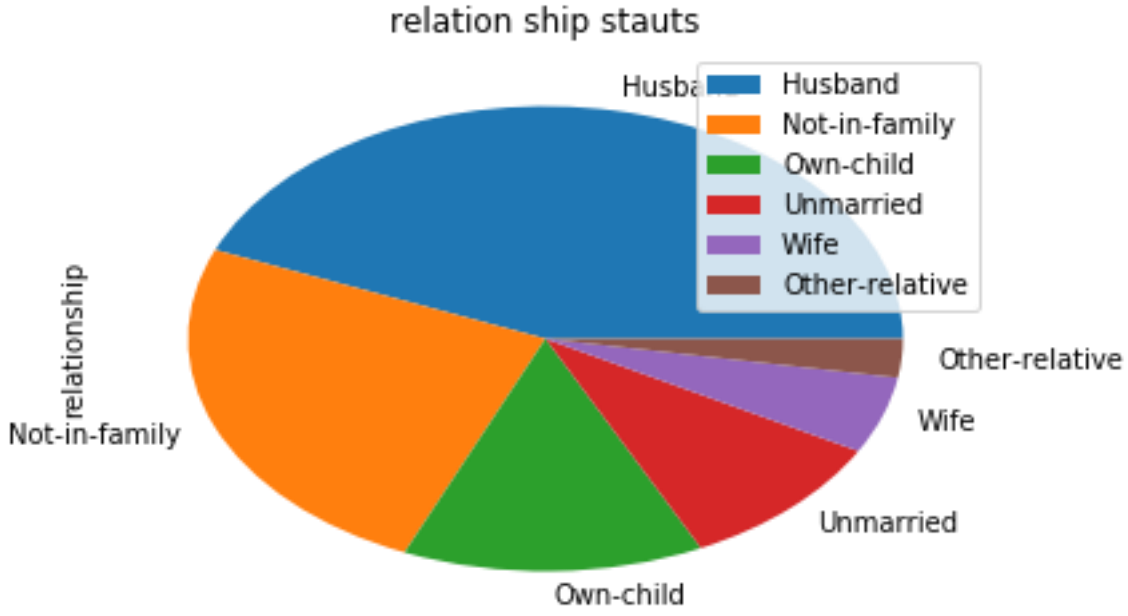
Marital Status:



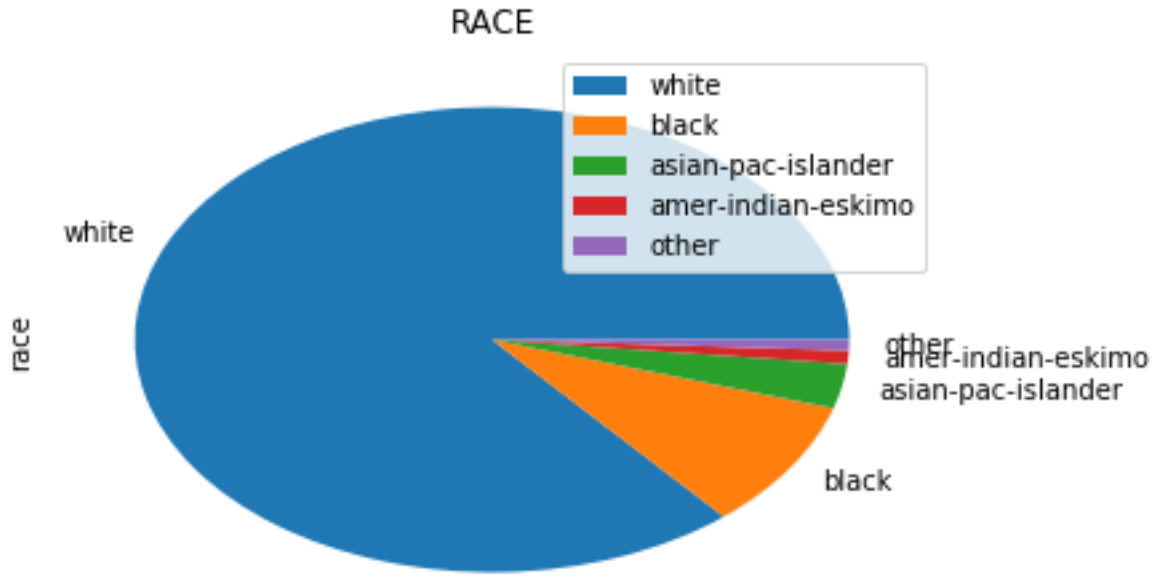
Occupation:



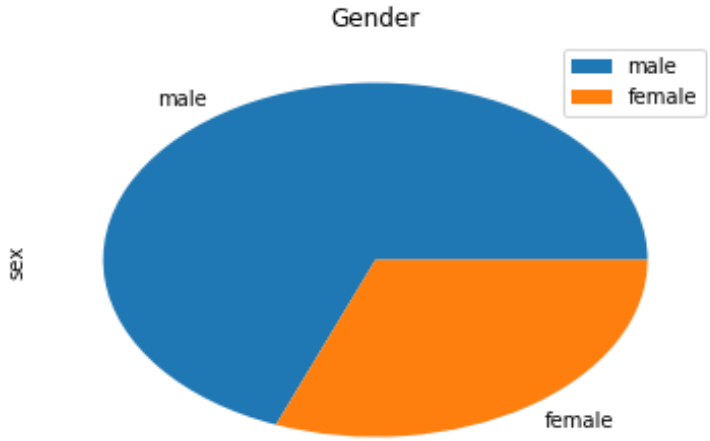
Relationship:



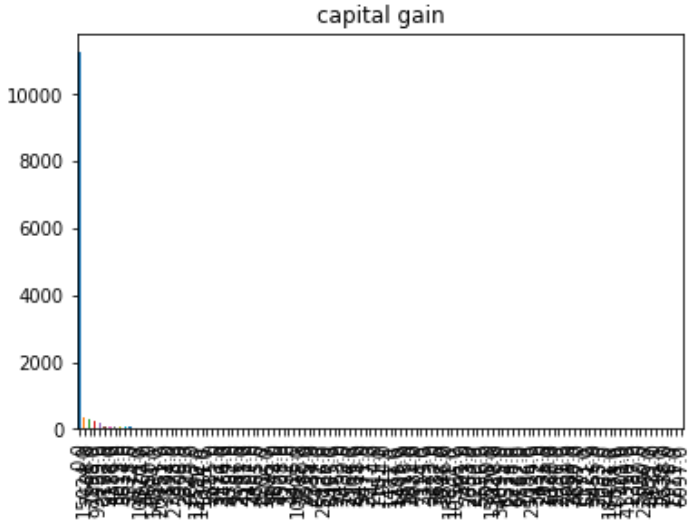
Race:



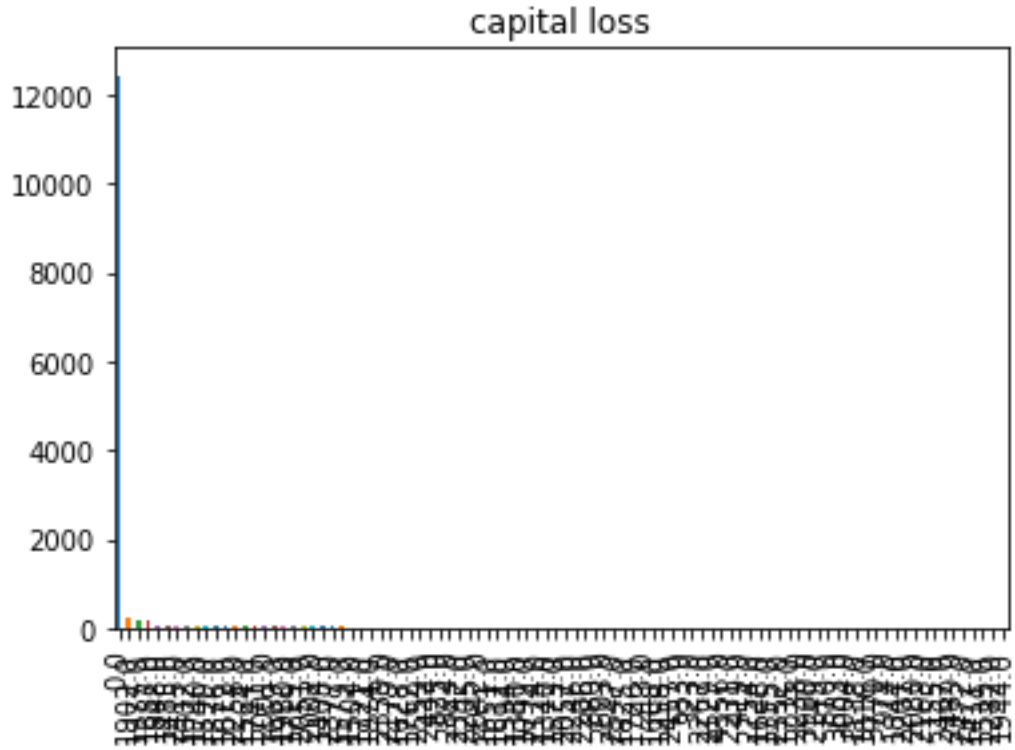
Sex:



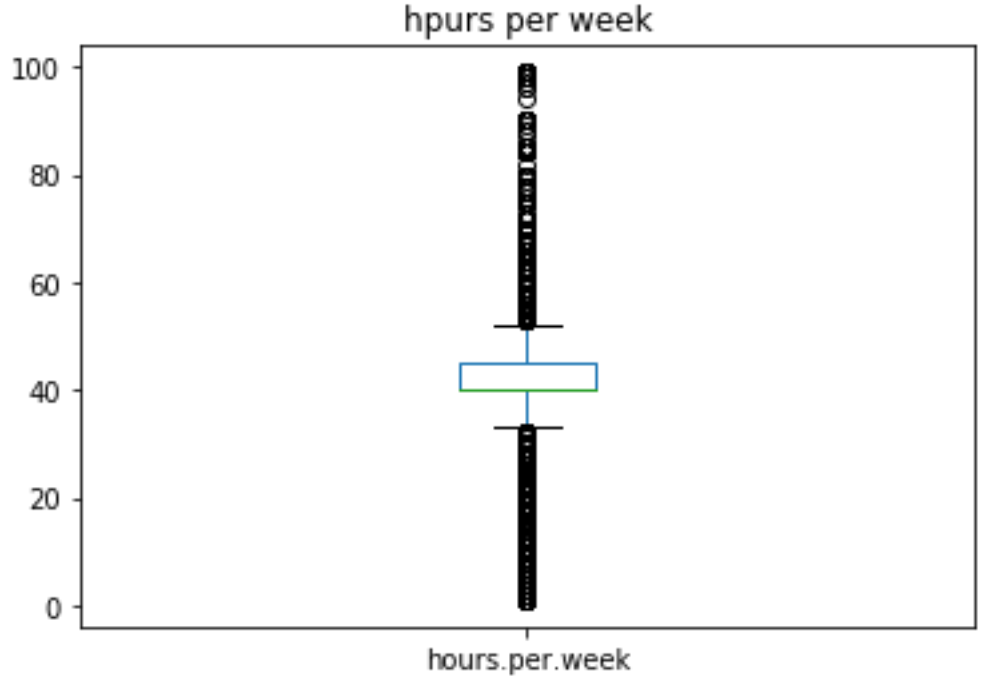
Capital Gain:



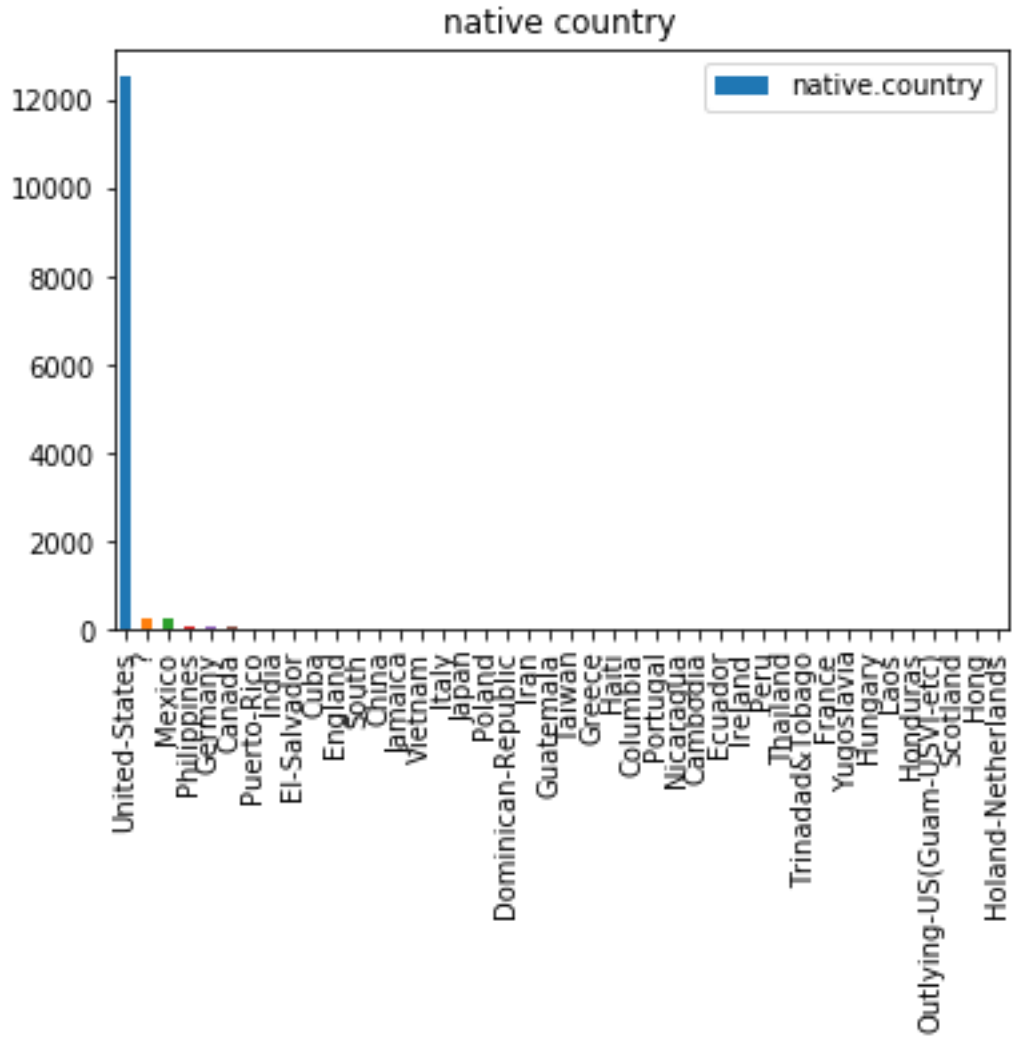
Capital Loss:



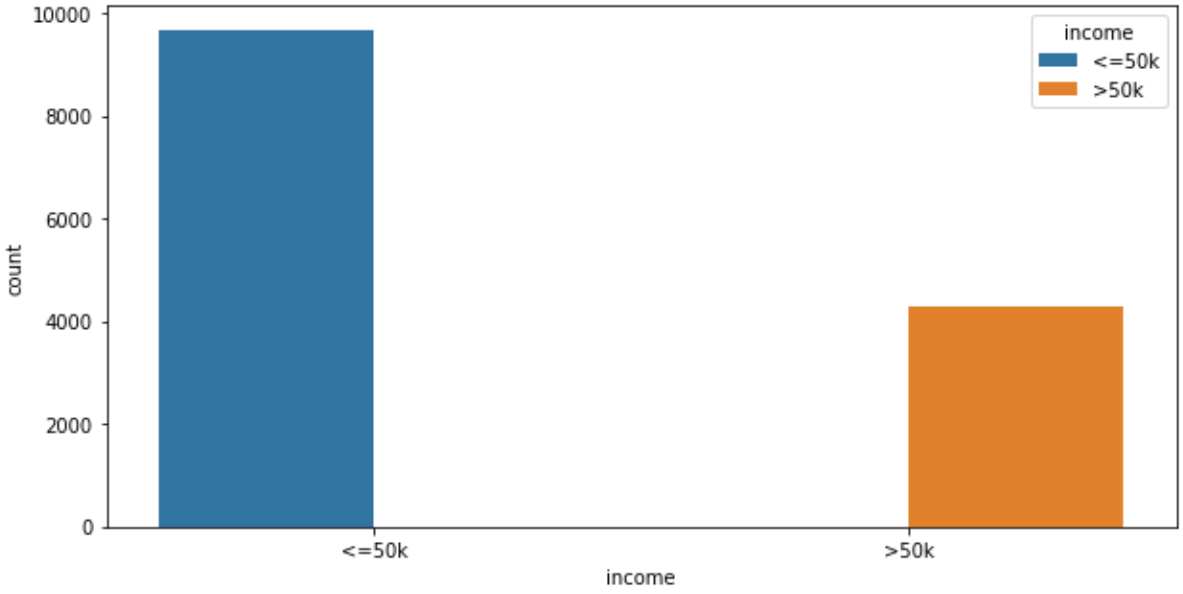
Hours per Week:



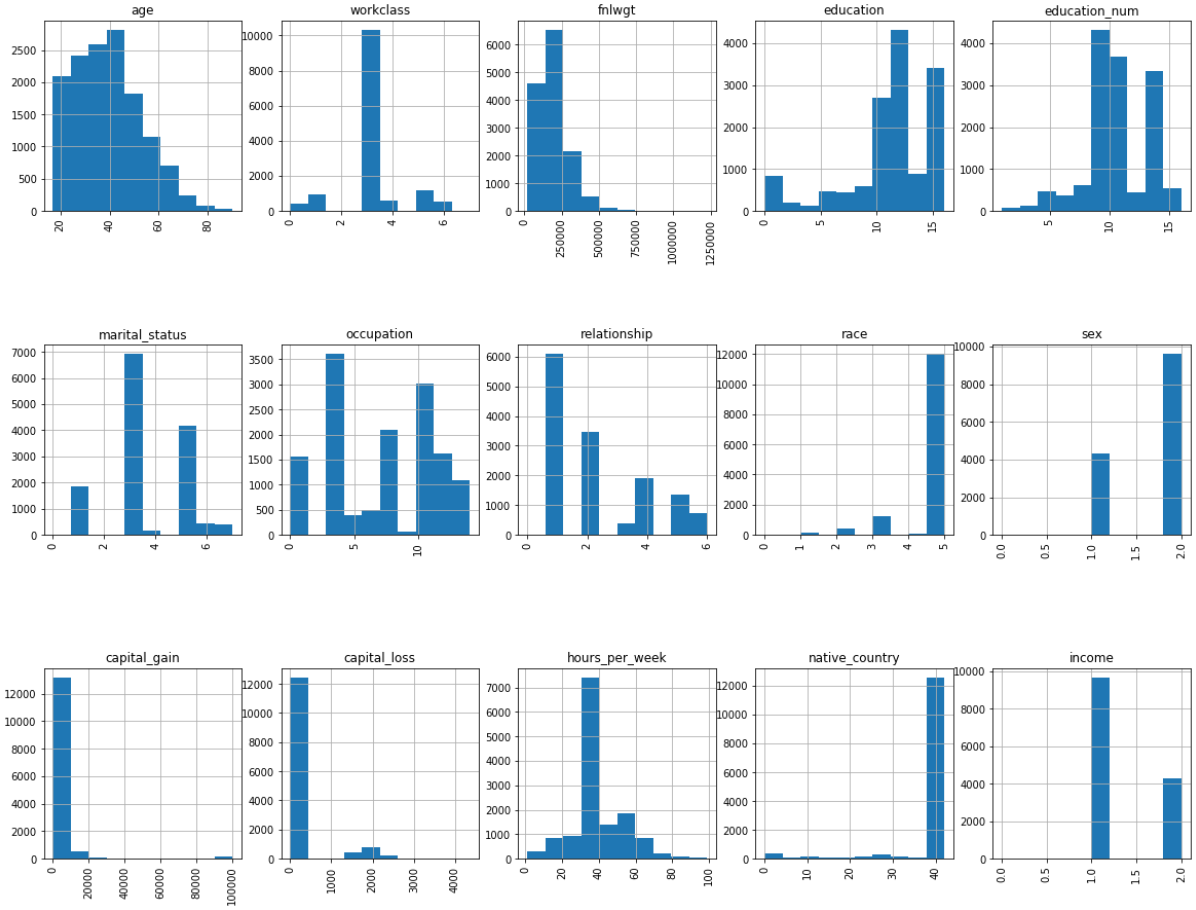
Native Country:


native country
native.country
12000
10000
8000
6000
4000
2000
0
United-States
?
Mexico
Philippines
Germany
Canada
Puerto-Rico
India
El-Salvador
Cuba
England
South
China
Jamaica
Vietnam
Italy
Japan
Poland
Dominican-Republic
Iran
Guatemala
Taiwan
Greece
Haiti
Columbia
Portugal
Nicaragua
Cambodia
Ecuador
Ireland
Peru
Thailand
Trinadad&Tobago
France
Yugoslavia
Hungary
Laos
Honduras
Outlying-US(Guam-USVI-etc)
Scotland
Hong
Holand-Netherlands

Income:

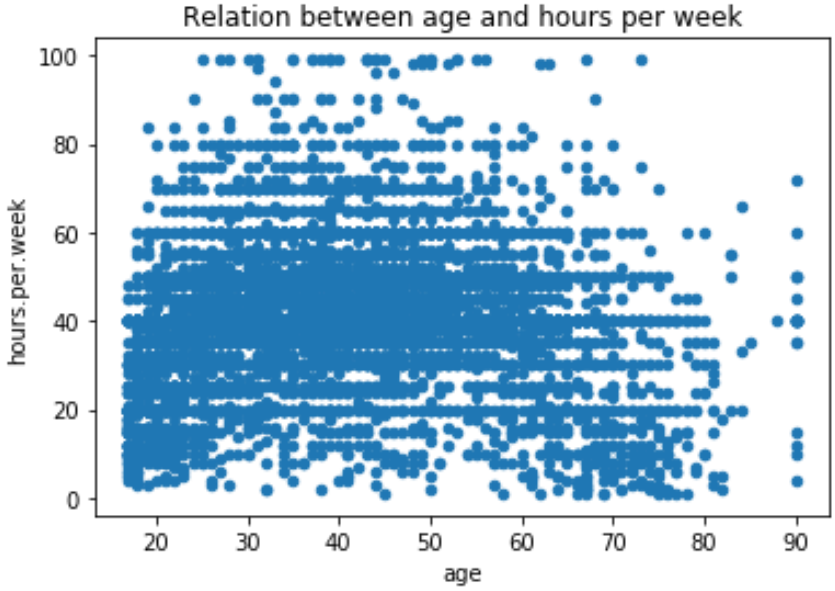

income
<=50k
>50k
10000
8000
6000
4000
2000
0
count
<=50k
>50k
income

Visualisation for every column in the data set:

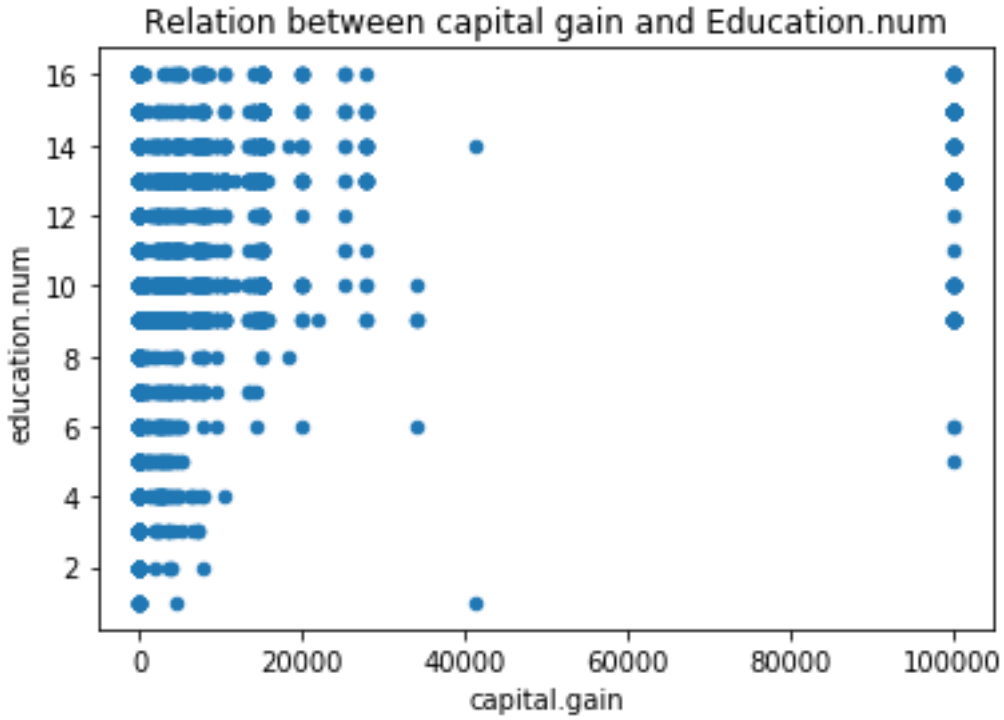


Relations:

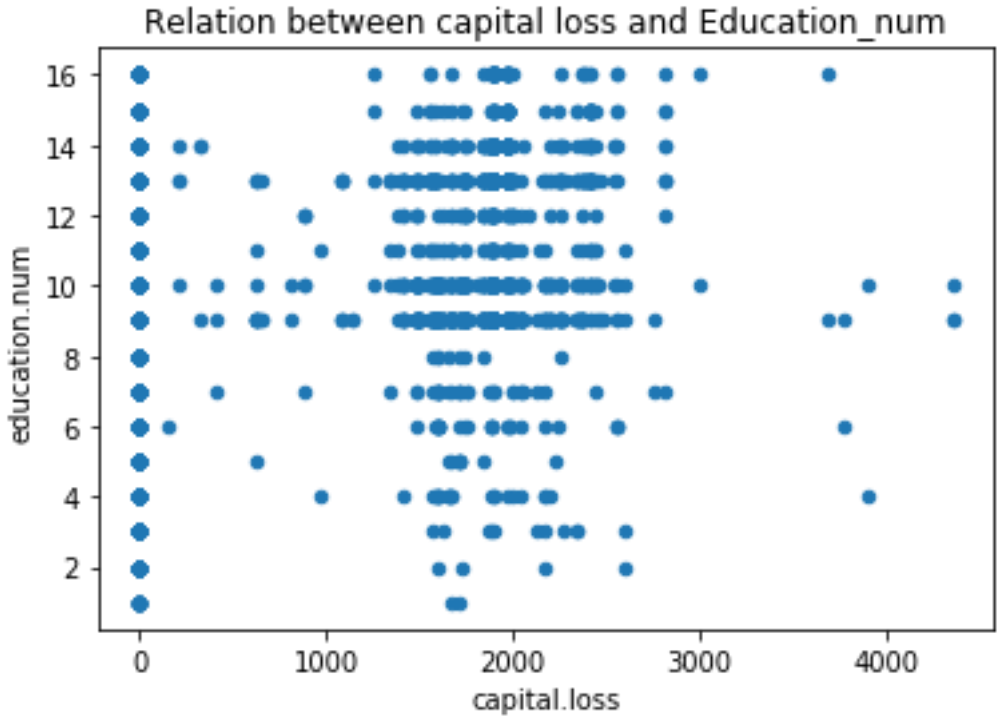
Age and Hours per week:



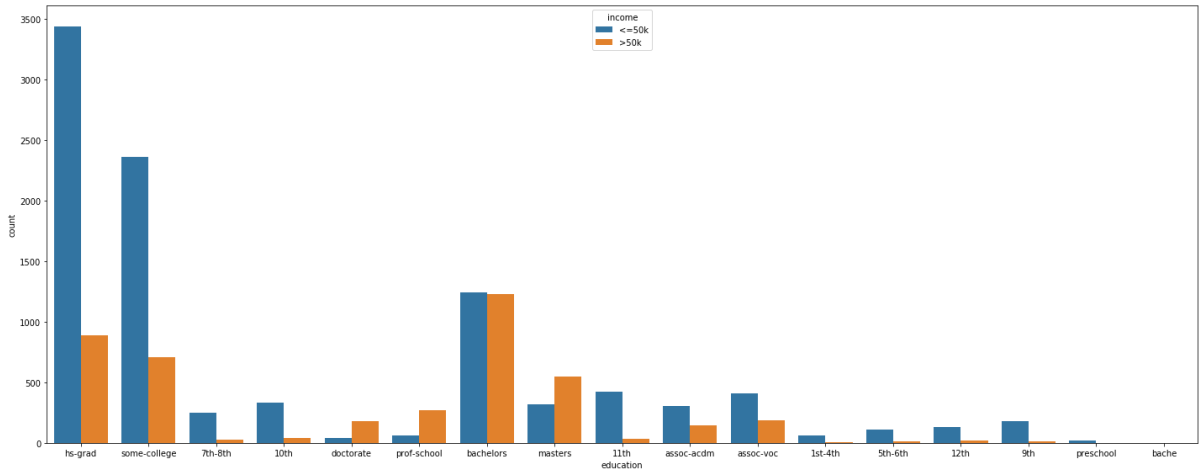
Capital Gain and Education Num:



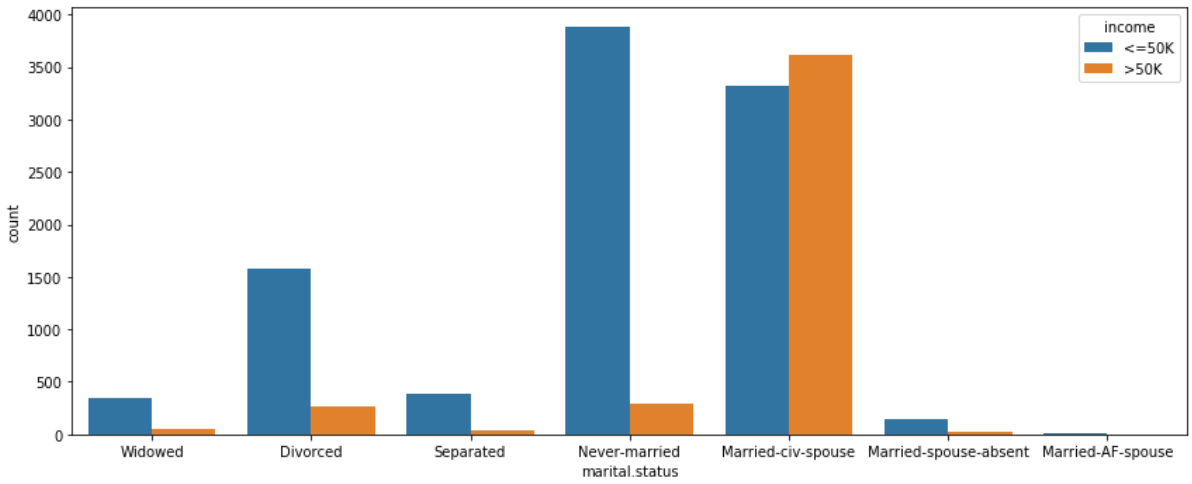
Capital loss and Education Num:



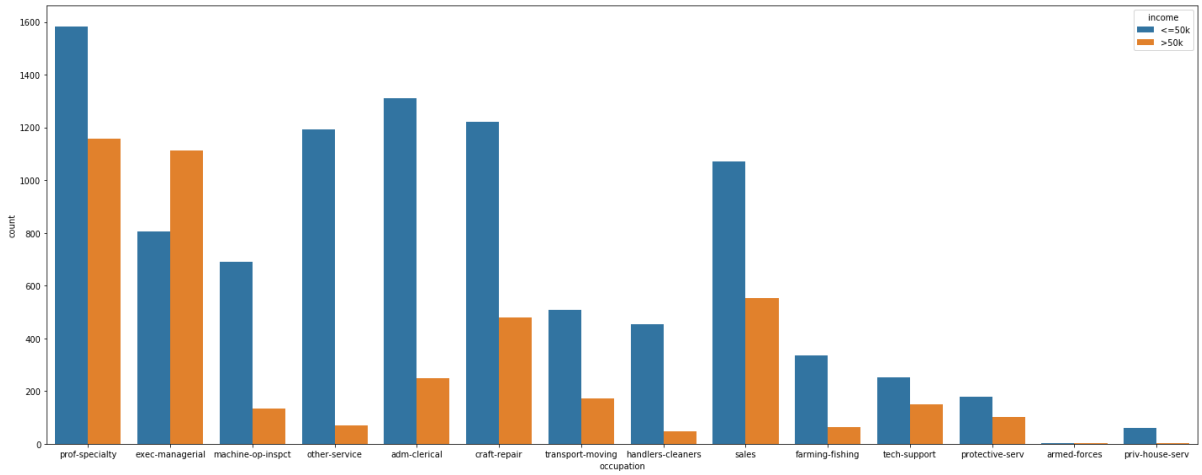
Education and Income:



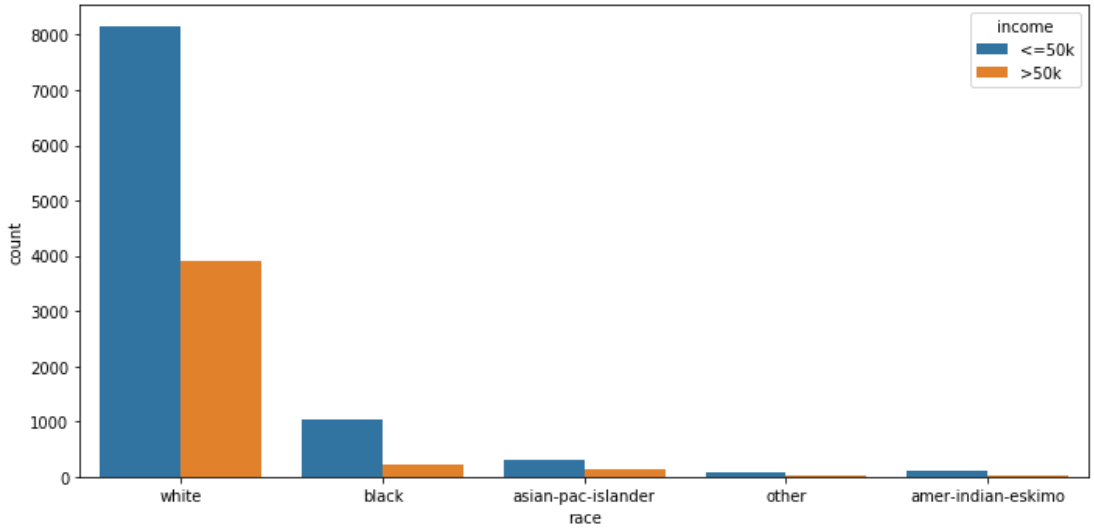
Marital Status and Income:



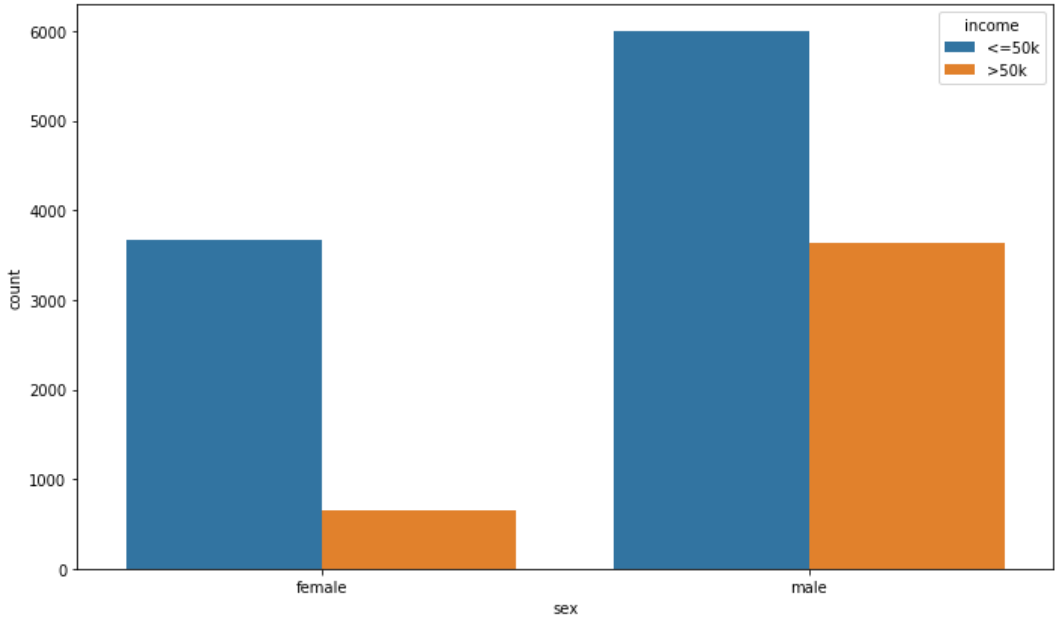
Occupation and Income:



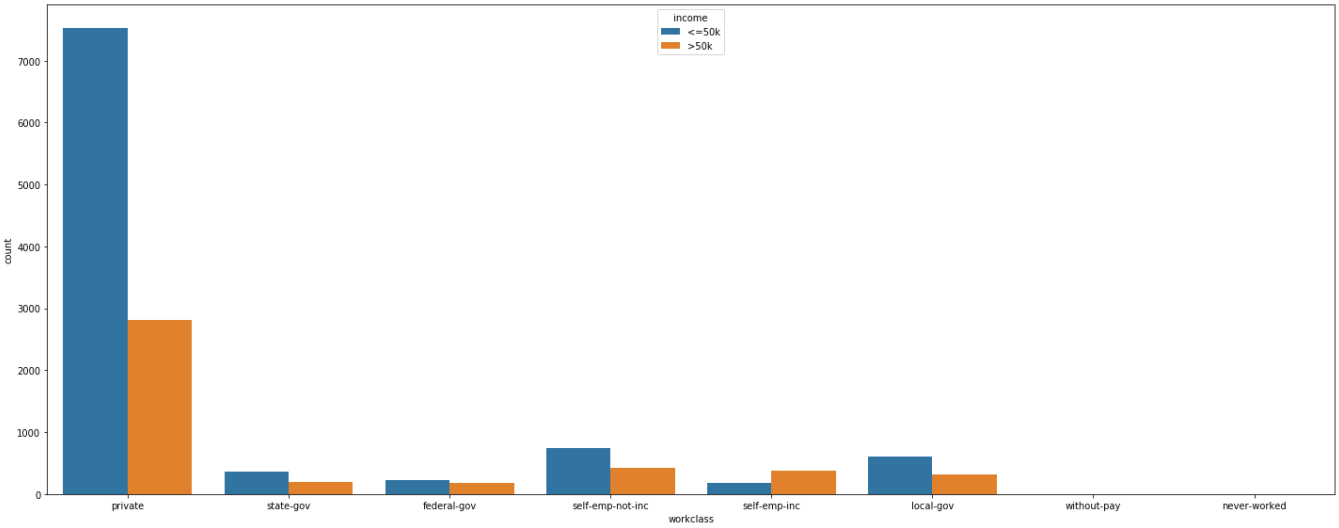
Race and Income:



Sex and Income:



Work Class and Income:



3.1 Modelling:

Modelling of the data is performed by splitting the dataset into two parts i.e training data and test data.

```
X = income[features].values
y = income[target].values.flatten()
```

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=0)
```

After splitting the data, we had applied Logistic Regression classifier with default parameters. Logistic Regression is a classification algorithm that is used to predict the probability of a categorical variable.

```python
def logistic_regression():

    # Instantiates and trains a logistic regression model using grid search
    # to find optimalhyperparameter values

    from sklearn.linear_model import LogisticRegression
    from sklearn.model_selection import GridSearchCV

    logreg = LogisticRegression()
    grid_values = {'penalty' : ['l1', 'l2'], 'C': [0.01, 0.1, 1, 10, 100]}

    grid_lr_rec = GridSearchCV(logreg, param_grid = grid_values, scoring = 'accuracy')
    grid_lr_rec.fit(X, y)

    return grid_lr_rec.best_estimator_
```

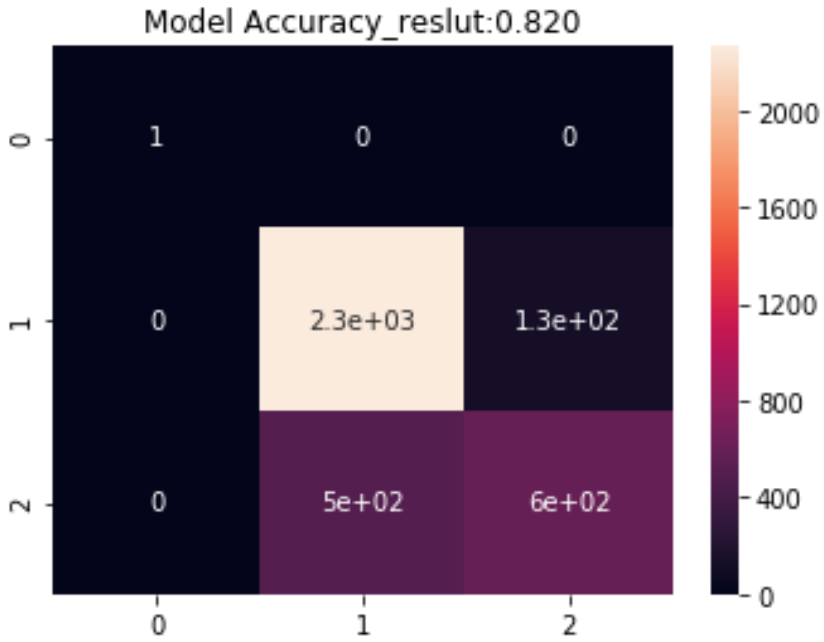
```python
logreg = logistic_regression()
```

```python
def evaluation(model):
    from sklearn.metrics import classification_report

    # This function gives provides verious evaluation metrics for the input model
    y_pred = model.predict(X_test)

    print(classification_report(y_test, y_pred, target_names = ['at most 50k', 'more than 50k']))
    print('Training data_Set Accuracy_result: {:.2f}'.format(model.score(X_train, y_train)))
    print('Testing data_Set Accuracy_result: {:.2f}'.format(model.score(X_test, y_test)))
```

After applying LR Algorithm, following is the evaluation report;

```
               precision    recall  f1-score   support

  at most 50k       1.00      1.00      1.00         1
more than 50k       0.82      0.94      0.88      2397

  avg / total       0.82      0.82      0.81      3491

Training data_Set Accuracy_result: 0.83
Testing data_Set Accuracy_result: 0.82
```

From the above figure we can see that we got an testing accuracy of 0.82% and also by observing the precision, recall ad f1 score we can see that there are more number of individuals lying in less than $50,000 threshold. A confusion matrix is visualized as follows,



From above confusion matrix, we can observe that the diagonal cells (2.3e+6e) indicates that there are more number of individuals lying in less than $50,000 threshold compared to other(1.3e+5e) as the cell 2.3e consists more number of individuals.

After LR algorithm we had applied Decision Tree with default parameters.

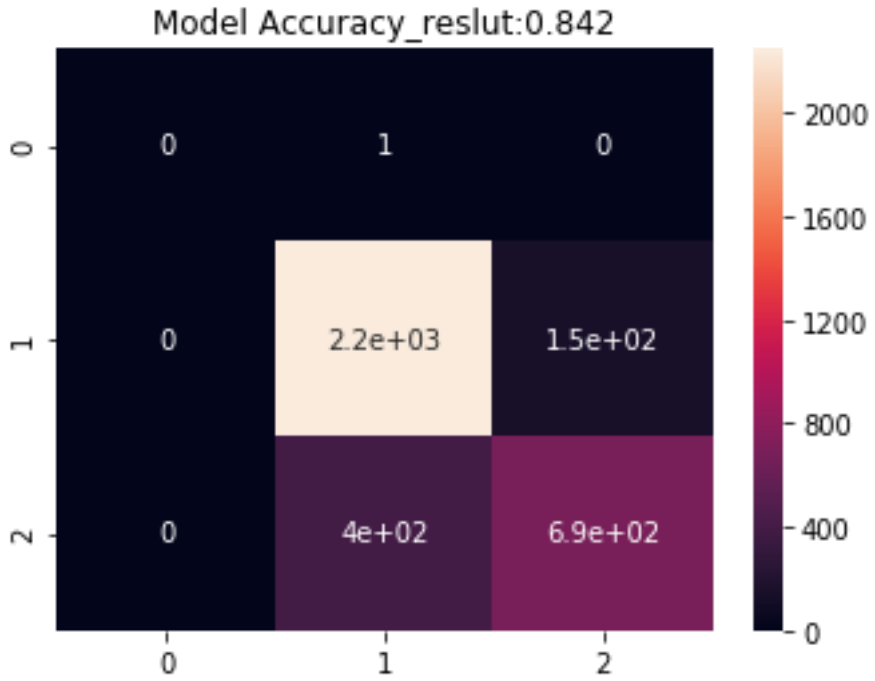
```
#Decision tree

def decision_tree():
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.model_selection import GridSearchCV

    grid_values = {'max_depth':np.arange(1,10), 'min_samples_leaf': np.arange(1,50,10)}
    clf = DecisionTreeClassifier()

    grid_dec_tree = GridSearchCV(clf, param_grid = grid_values)
    grid_dec_tree.fit(X, y)

    return grid_dec_tree.best_estimator_
```

Then we had received an evaluation report after applying decision tree algorithm.

	precision	recall	f1-score	support
at most 50k	0.00	0.00	0.00	1
more than 50k	0.85	0.94	0.89	2397
avg / total	0.84	0.84	0.84	3491

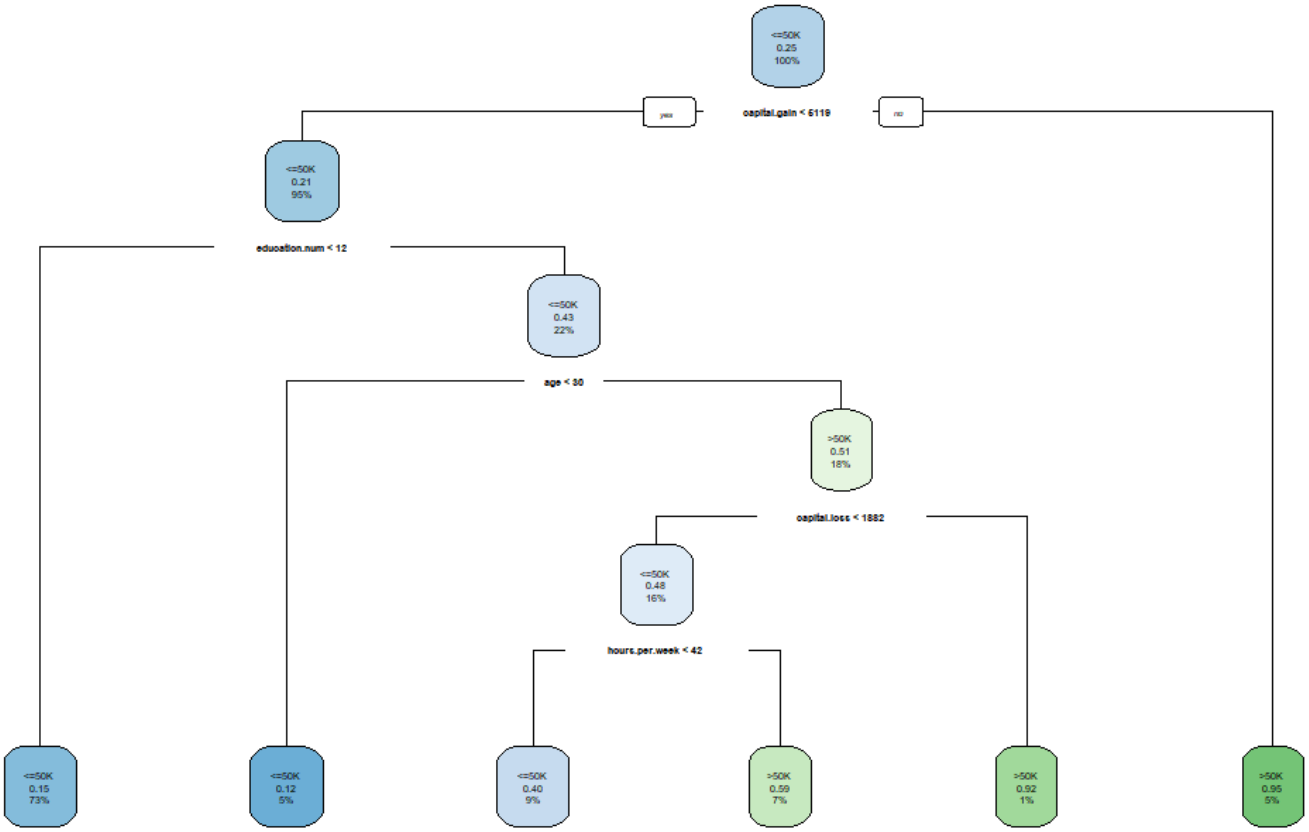
```
Training data_Set Accuracy_result: 0.84
Testing data_Set Accuracy_result: 0.84
```

From above figure we can say that we got an accuracy of 0.84% which is higher than LR algorithm and also by observing the precision, recall ad f1 score we can see that there are more number of individuals lying in less than $50,000 threshold.. The confusion matrix is displayed as follows



From above confusion matrix, we can observe that the diagonal cells (2.2e+6.9e) indicates that there are more number of individuals lying in less than $50,000 threshold compared to other(1.5e+4e) as the cell 2.2e consists more number of individuals compares to others.

A decision tree diagram is displayed as follows ,



In the above diagram, we can observe that there more number of people earning less than US$50,000. So,we can confirm that there are more number of individuals earning less than US$50,000 per annum.

We also applied RandomForest Classifier.

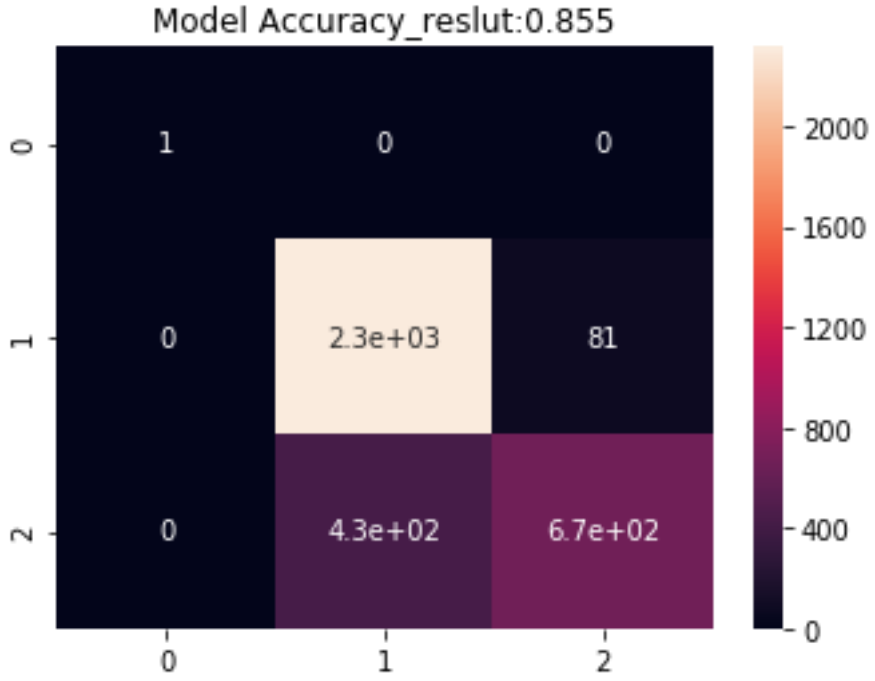
```
#random Forest Classifier
def random_forest():
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.model_selection import GridSearchCV

    clf = RandomForestClassifier(random_state=0)
    grid_values = {'max_depth': np.arange(1,11,2), 'max_features': np.arange(1,11,2)}

    grid_clf = GridSearchCV(clf, param_grid = grid_values)
    grid_clf.fit(X, y)


    return grid_clf.best_estimator_
```

Then the evaluation results are as follows

	precision	recall	f1-score	support
at most 50k	1.00	1.00	1.00	1
more than 50k	0.84	0.97	0.90	2397
avg / total	0.86	0.85	0.85	3491

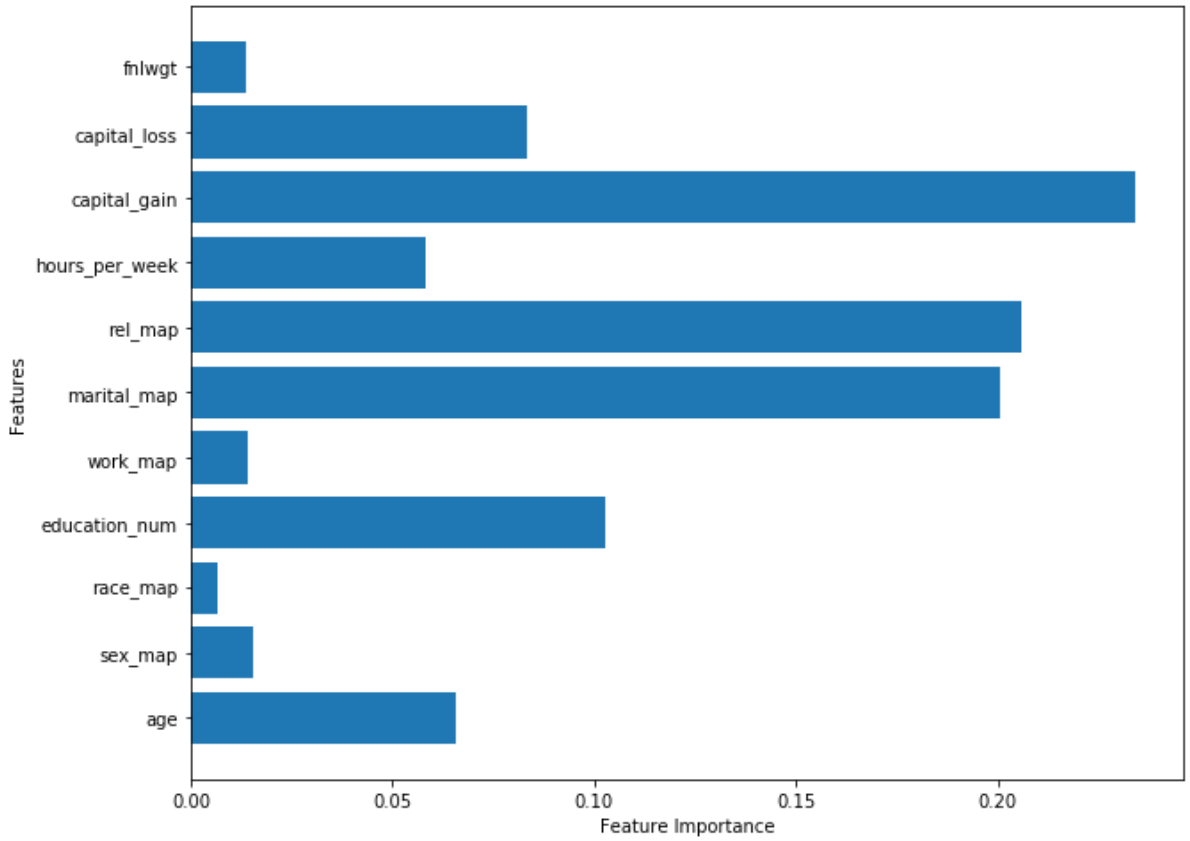
```
Training data_Set Accuracy_result: 0.86
Testing data_Set Accuracy_result: 0.85
```

From above results the test accuracy was 0.85% which is greater than Decision tree and LR classifier. **B**y observing the precision, recall ad f1 score we can see that there are more number of individuals lying in less than $50,000 threshold.. The confusion matrix is displayed as follows



From above confusion matrix, we can observe that the diagonal cells (2.3e+6.7e) indicates that there are more number of individuals lying in less than $50,000 threshold compared to other(181+4.3e) as the cell 2.3e consists more number of individuals compares to others.

From Random Forest Classifier we were able to find out some important features of this data. It is visualized as follows,



4. Conclusion

After doing the classification, Decision Tree has the highest test accuracy of 0.84% compared to LR algorithm of 0.82% and we can conclude that there are more number of individuals earning less than US$50,000 per annum.

5. References

UCI Archive, 2011, *http://archive.ics.uci.edu/ml/machine---learning---databases/adult/adult.names*