Neural Network Assignment Report:

Features:

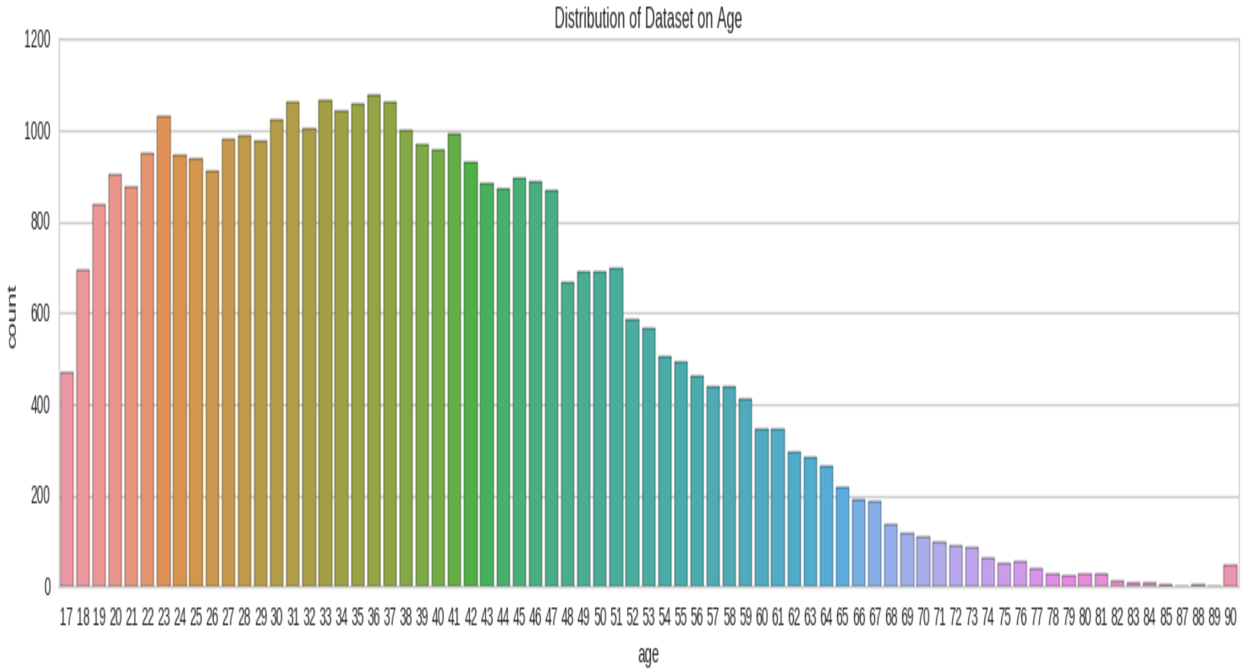
The given data set has 16 columns.
Id,age,workclass,fnlwgt,education,education-num,marital-status,occupation, relationship,race, sex,capital-gain,capital-loss,hours-per-week,native-country,salary.

Numerical features are:
Age,fnlwgt,education-num,capital-gain,capital-loss,hours-per-week,salary.

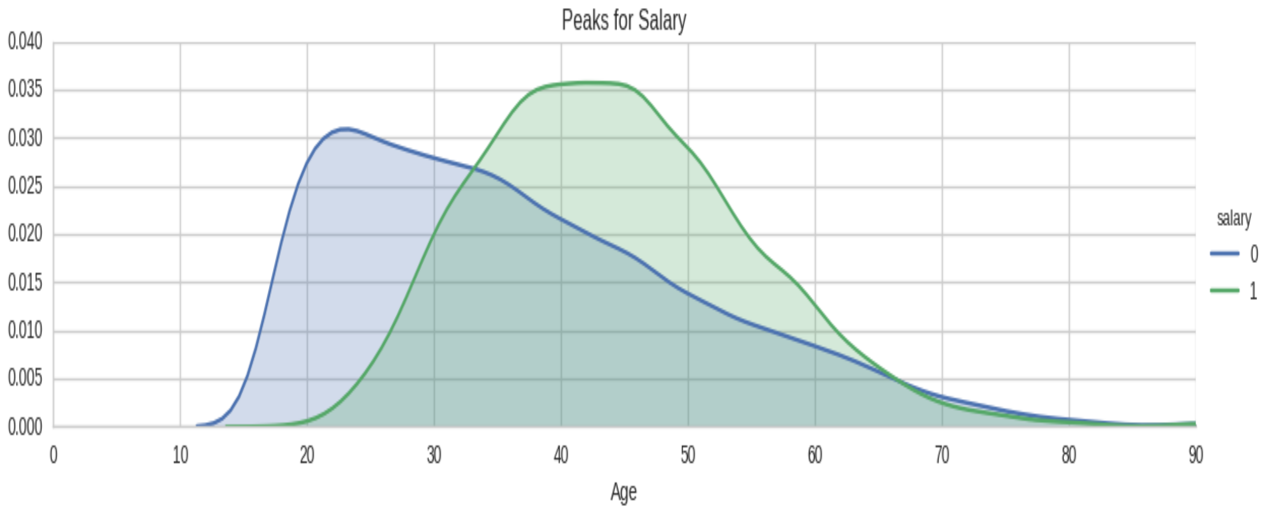
Categorical features are:
Work class,education,marital-status,occupation,relationship,race,sex,native-country.

Age:

The range of weight is 17-90.Most of the data is from age 19-47.
Distribution of age is shown below:

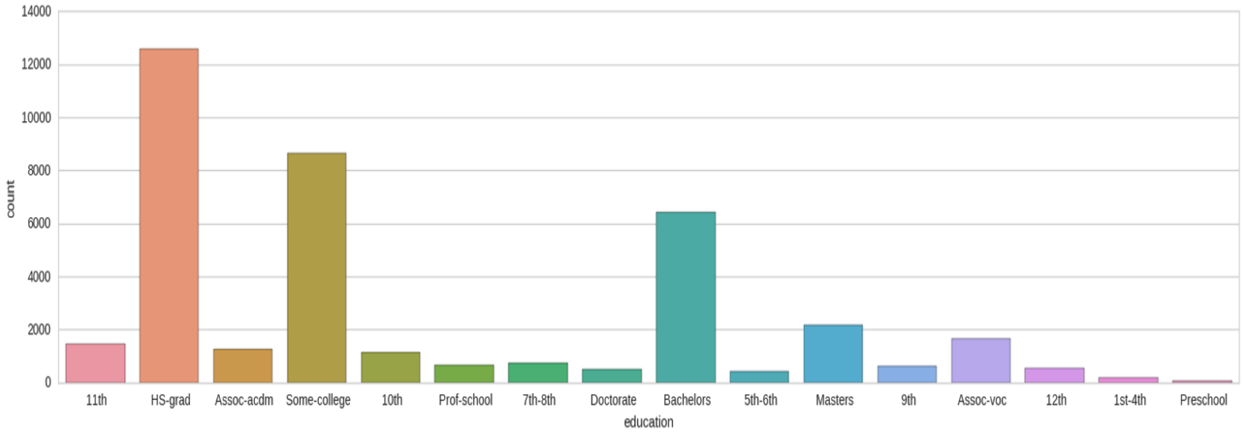


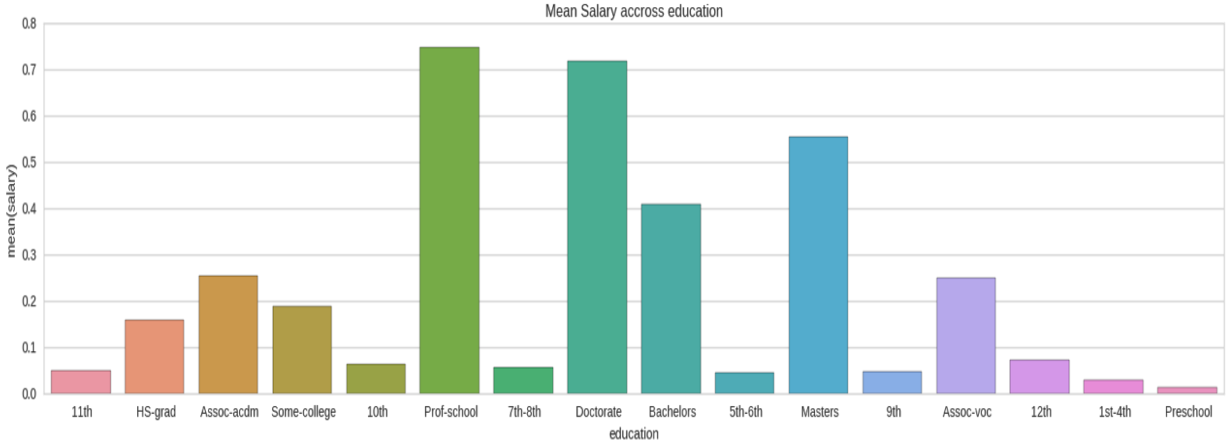
Most of the people of the age 22-24 get salary < 50k and most of the people of the age 42-45 get salary >=50k.



Education :

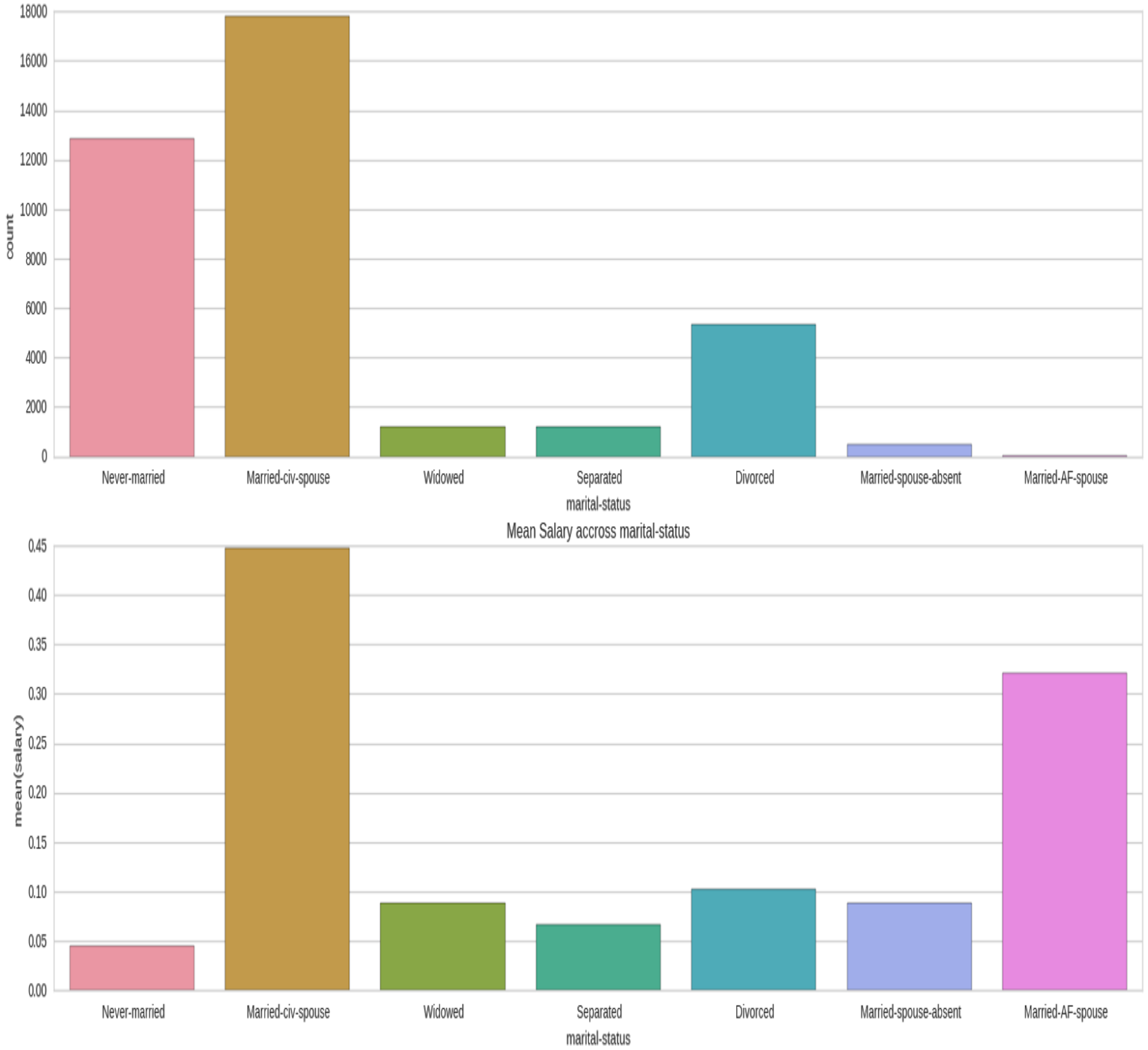
Most of the dataset is of High School grads, Bachelors and College people *Masters,* Doctorate and Professors in School have high probability of getting salary > 50k.





Marital-status:

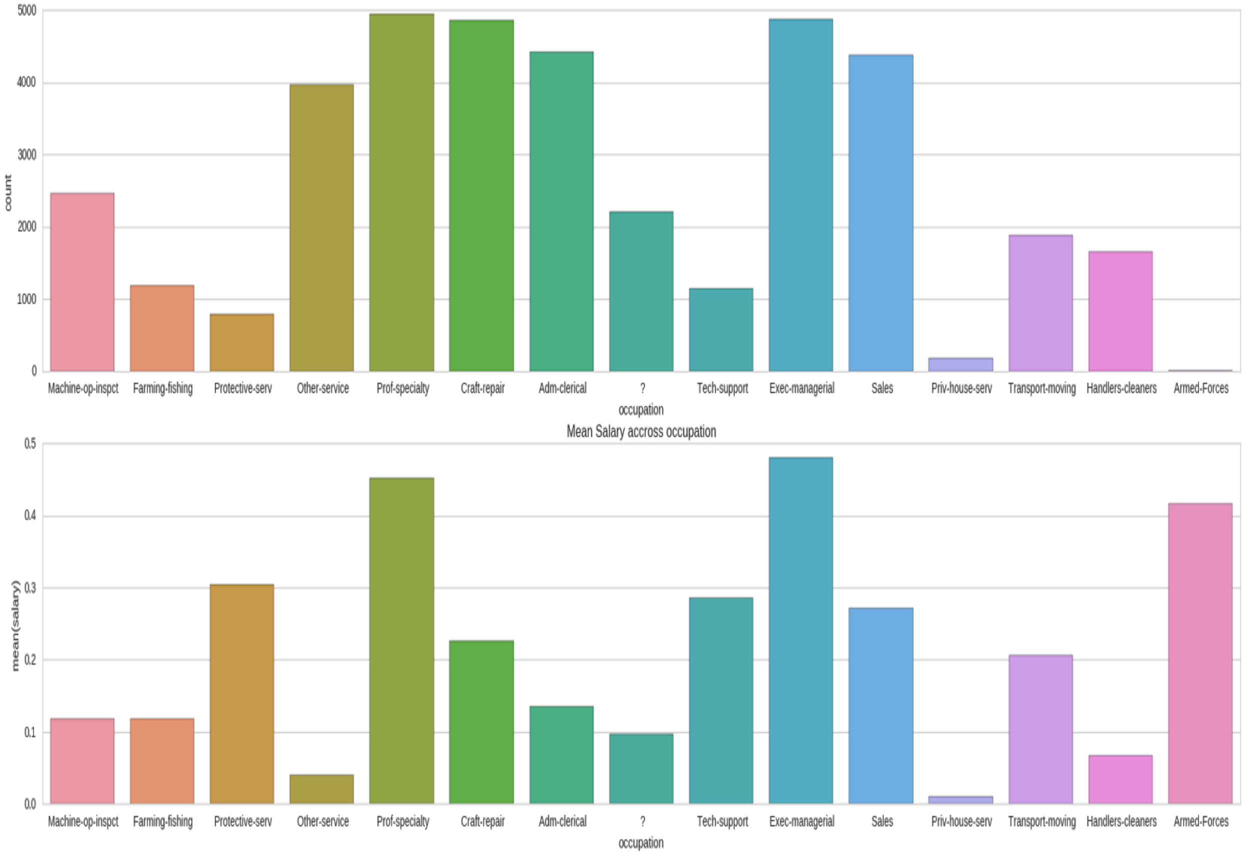
Maximum dataset belong to “Married-civ-spouse” category.
Mean salary of “Married-civ-spouse” category is highest.And for “never-married” category is lowest.



Occupation:

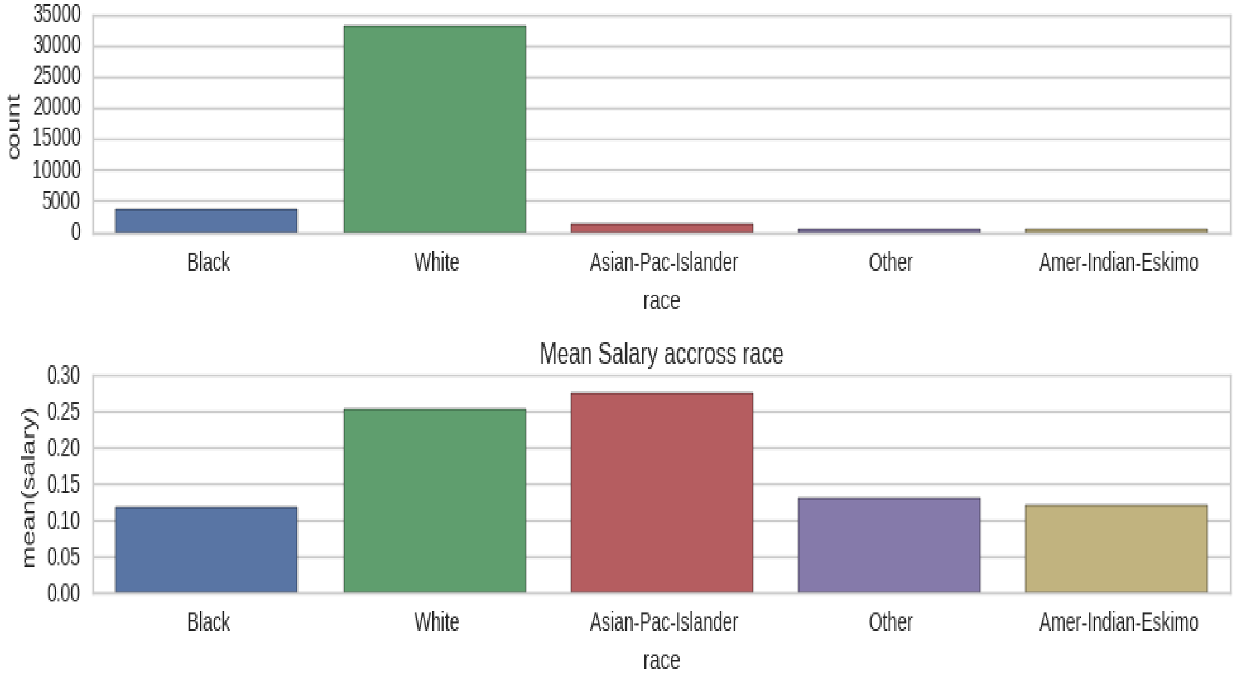
Maximum mean salary are for the category “prof-speciality” and “exec-managerial”.

Minimum mean salary belongs to the catagory “priv-house-serv”.



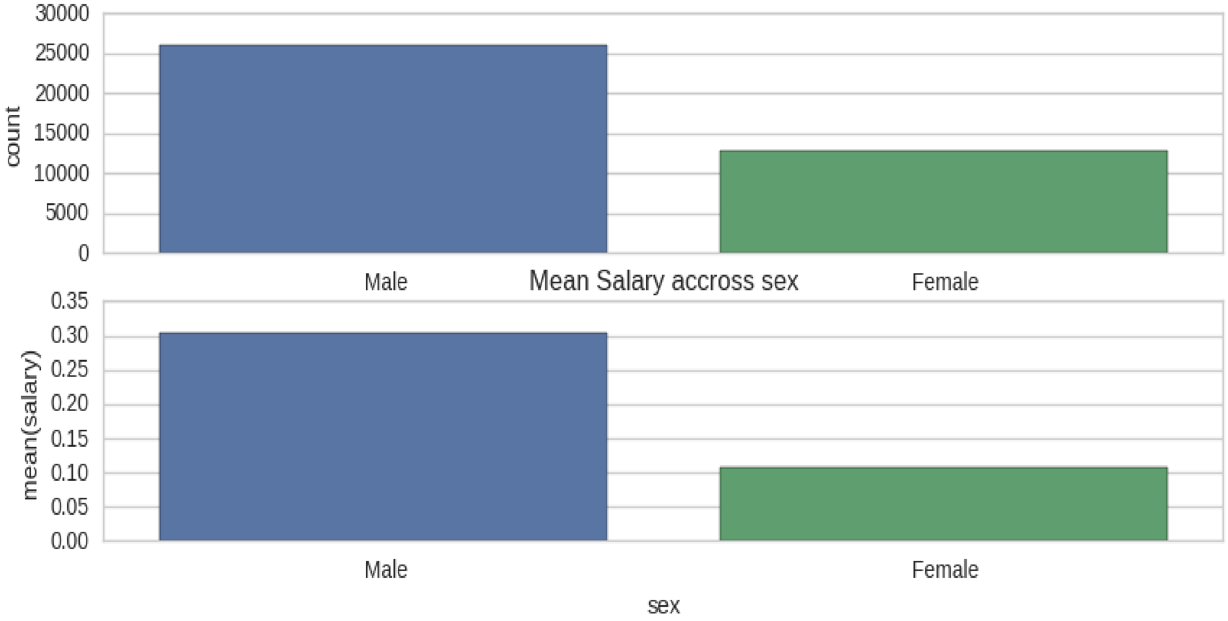
Race:

Most dataset is from “White” category of race.
Maximum mean salary belongs to the “White” and “Asian-pac-Islander-race”.



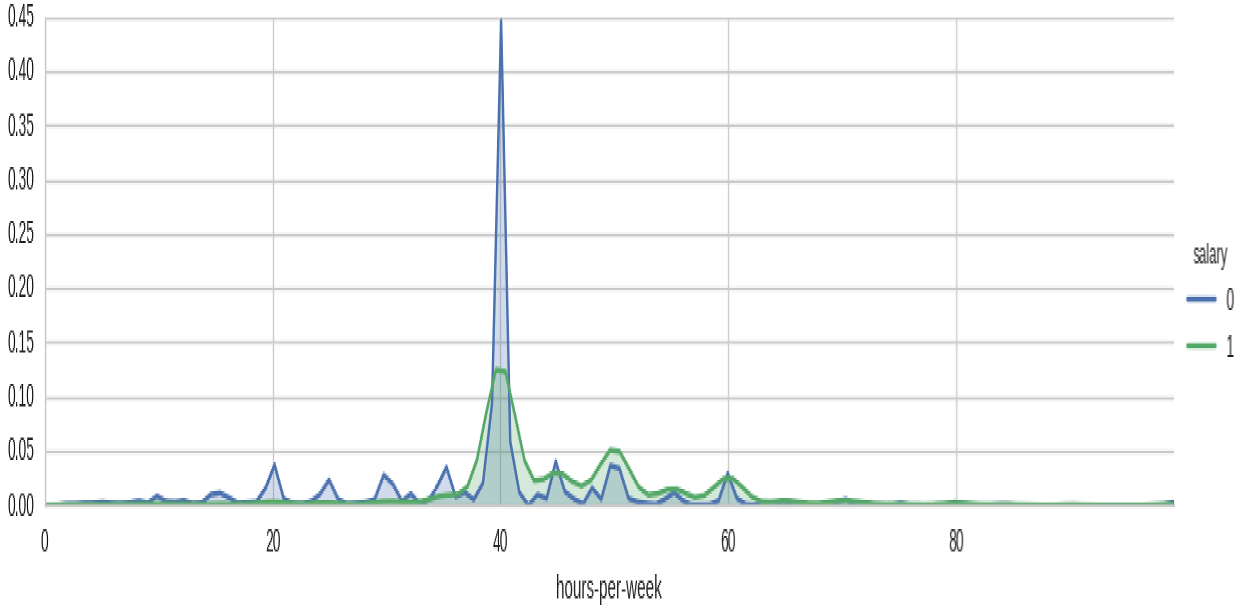
Sex:

Most of the dataset is from the “Male” catagory for sex.And “Male” catagory has higher mean salary than “Female”.



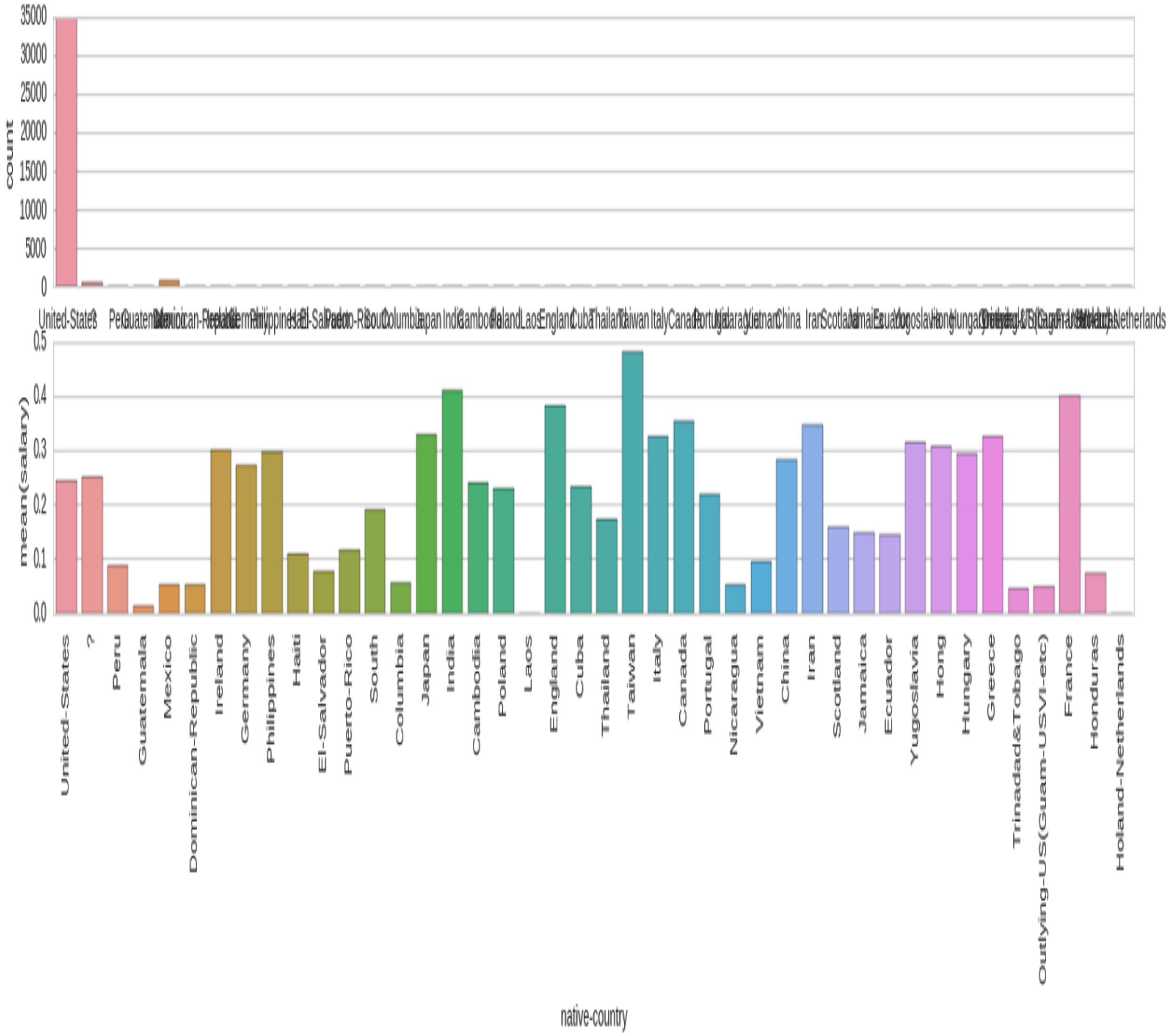
Hours-per-week:

Maximum peple with <50K salary work for 40+-3 hours.

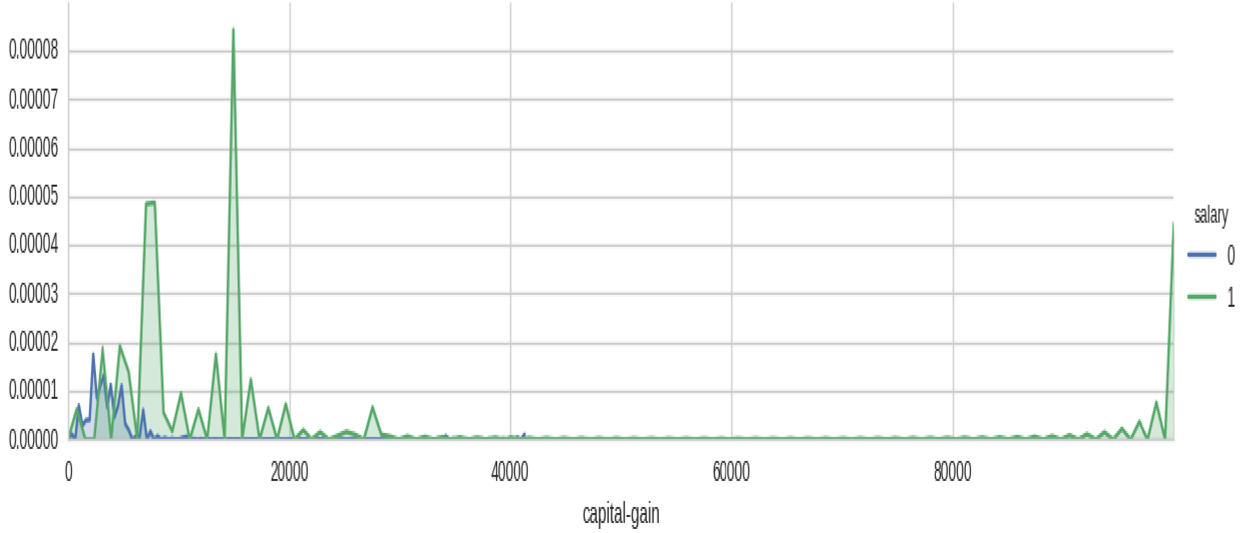


Native-country:

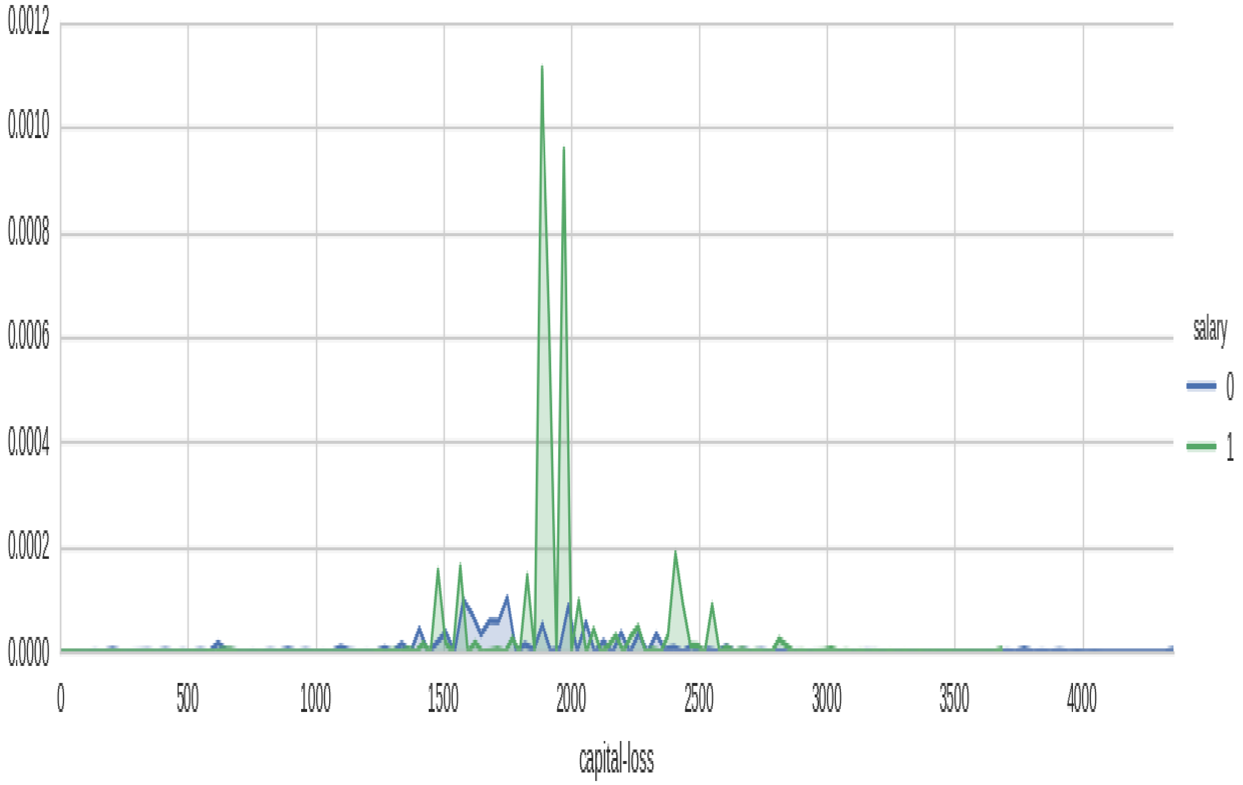
More than 90% data belong to “USA” for native-country catagory.So I removed native-country as it does not give too much value for the prediction.



Capital-gain:



Capital-loss:



Working of code:

train_net.py : This is the main file for neural-network implementation.

NN class: This is the class for neural network.NN constructor will take number of neurons in each layer(1 input,1 hidden and 1 output) and it will initialize activation matrices and weight matrices accordingly.

def update() : Performs forward propagation.

def backPropagate(): Performs backward propagation.

def train(): Trains the data set using update() and backPropagate() method for a given number of iterations.

def getData() : Function which reads train data from train.csv and sends to train() for training.

def sigmoid() : Takes an input and returns a value after applying tanh() function in it.

def dsigmoid() : Returns the value after applying derivative of tanh() function on given input.

def normalize() : Function to normalize the data.

def demo(): Function which will execute first.The function will create a neural network using NN class.The it will get the data for training using getData() method.It will the train the neural network using NN.train() method.After the training is complete it will save the final weight matrices(NN.wi and NN.wo) for testing in weights.txt file.

Comparison:

The Following methods I have used to compare the performance with Neural network:

1. Logistic Regression
2. Gaussian naive bayes
3. svm

1.The Logistic regression is giving AUC error on kaggle as **0.80162.**

2.Gaussian Naive Bayes is giving AUC error on kaggle as **0.73343.**

3.svm is giving AUC error on kaggle as **0.72359.**

Neural network implementation(Trained on 20000 data) is giving AUC error on kaggle **0.81258.**