|  |  |  |
| --- | --- | --- |
| Evaluation Project | March 24  2024 | |
| **Spandan Ray Choudhury / Batch no:DS2310** | | Data Analysis |

**PROBLEM STATEMENT:**

Earning money and work life balance is very important task to do. We have collected a data set where we have divided the earning in 02 parts one is greater than 50k and other one is less than 50k, by using existing data set we need to analyze the working hour per week depending upon other variable, available in data set.

**Introduction of column::**

* **Age**:
  + The age of an individual
  + Integer data type greater than 0
* **Work class**:
  + Employment status of an individual
  + Type of organization in he is working.
* **Final weight(Fnlwght):**
  + Final weight is the number of people the census believes the entry represents.
* **Education:**
  + The highest level of education achieved by an individual.
* **Education\_­num:** 
  + The highest level of education achieved in numerical form.
* **Marital ­status:** 
  + Marital status of an individual.
* **Occupation:** 
  + Occupation type in which field he/she is working.
* **Relationship:** 
  + Represents what this individual is relative to others.
* **Race:** 
  + Descriptions of an individual’s race
* **sex:**
  + The biological sex of the individual
* **Capital ­gain:**
  + Capital gains for an individual, the initial investment have been increased.
* **Capital ­loss:** 
  + Capital loss for an individual, the initial investment has been decrease.
* **Native ­country:** 
  + country of origin for an individual (from where he belongs)
* **Earning**:
  + The earning has been divided into two parts one is below 50k (per annum earning is lower than 50k) another one is above 50k (per annum earning is greater than 50K)
* **Working Hour per week:**
  + **We need to calculate the working hour per week to go above if an individual needs to cross 50k mark depending upon the criteria mentioned above.**

**EXPLANATION OF DATA SET**

**We have collected data from various sources. The data set structure is as mentioned below-**

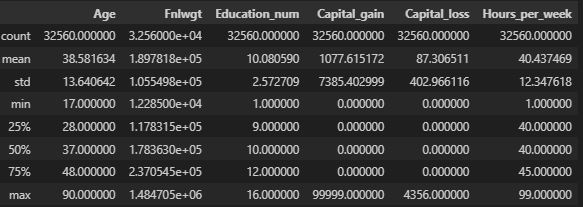
**Data structure and distribution**

|  |  |  |
| --- | --- | --- |
| Shape of Data | Null value | Type of value |
| 32560 Rows X 15 Columns | **0 – There are no null value in data set** | **Numerical and Categorical both values present** |

**Value type in columns:**

|  |  |  |
| --- | --- | --- |
| *Column name* | *Value\_type* | *Type of values(count)* |
| *Age* | ***Numerical*** | ***73*** |
| *Work Class* | ***Catagorical*** | ***9*** |
| *Fnlwgt* | ***Numerical*** | ***21647*** |
| *Education* | ***Categorical*** | ***16*** |
| *Education\_num* | ***Numarical*** | ***16*** |
| *Marital Status* | ***Categorical*** | ***7*** |
| *Occupation* | ***Categorical*** | ***15*** |
| *Relationship* | ***Categorical*** | ***6*** |
| *Race* | ***Categorical*** | ***5*** |
| *Sex* | ***Categorical*** | ***2*** |
| *Capital Gain* | ***Numarical*** | ***119*** |
| *Capital Loss* | ***Numerical*** | ***92*** |
| *Hours Per Week* | ***Numerical*** | ***94*** |
| *Native Country* | ***Categorical*** | ***41*** |
| *Income* | ***Categorical*** | ***2*** |

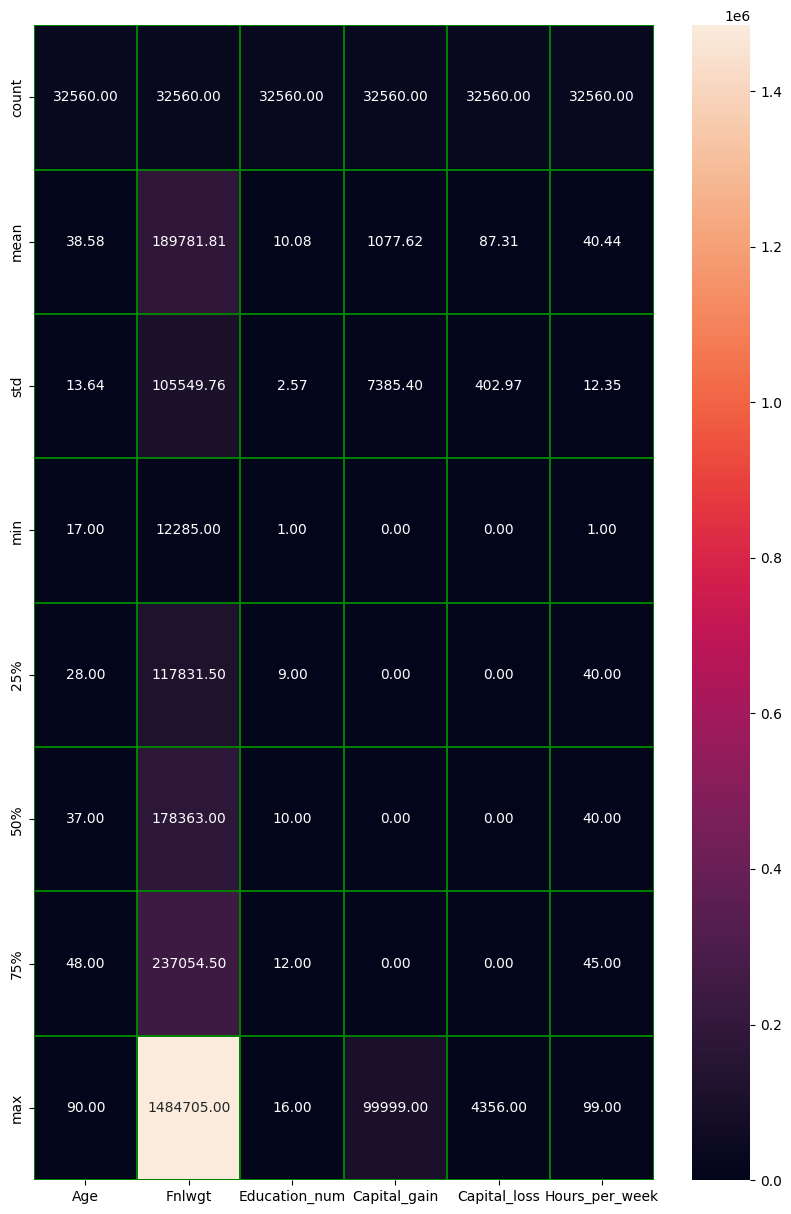
**Distribution and Description of Numerical data in the data set:**

****

**Fig:01(data description)**

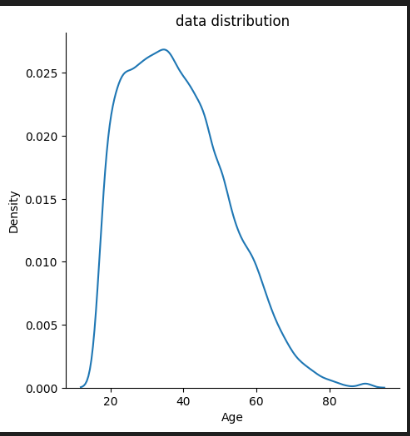
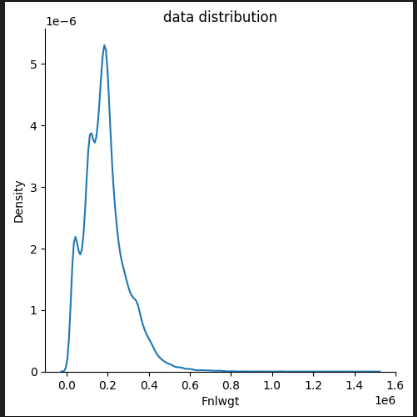
|  |
| --- |
| **Interpretation**   * **The mean value and median value is almost same of Age, Fnlwt, Education\_num and Hours\_per\_week columns. The data of this column is more or less normally distributed.** * **In capital gain and capital loss mean>median it means the data is right skewed.** * **As count= total no of rows then there is no presence of Null value.** |

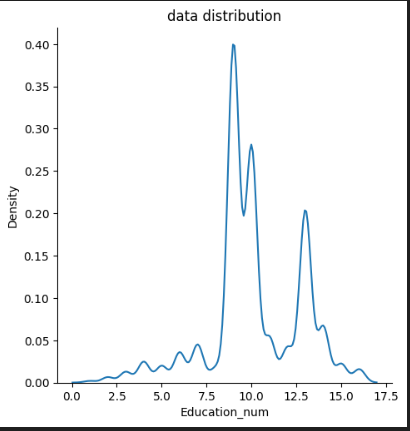
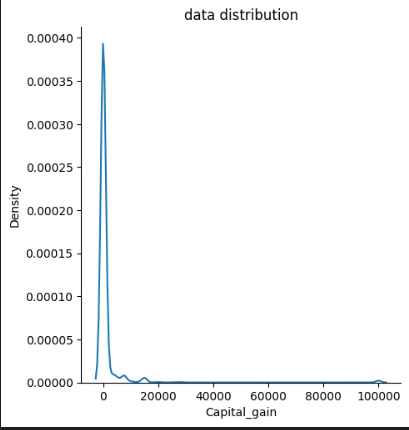
**DATA DESCRIPTION HEAT MAP OF NUMERICAL DATA:**

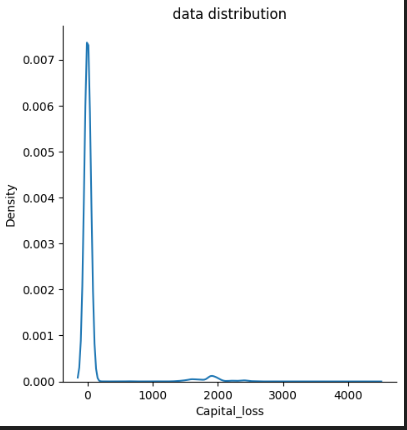
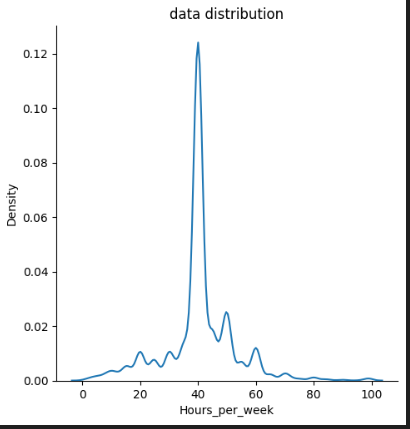
****

**Fig:02:heatmap of data description**

**GRAPHICAL DISTRIBUTION OF NUMARIC DATA:**

**** ****

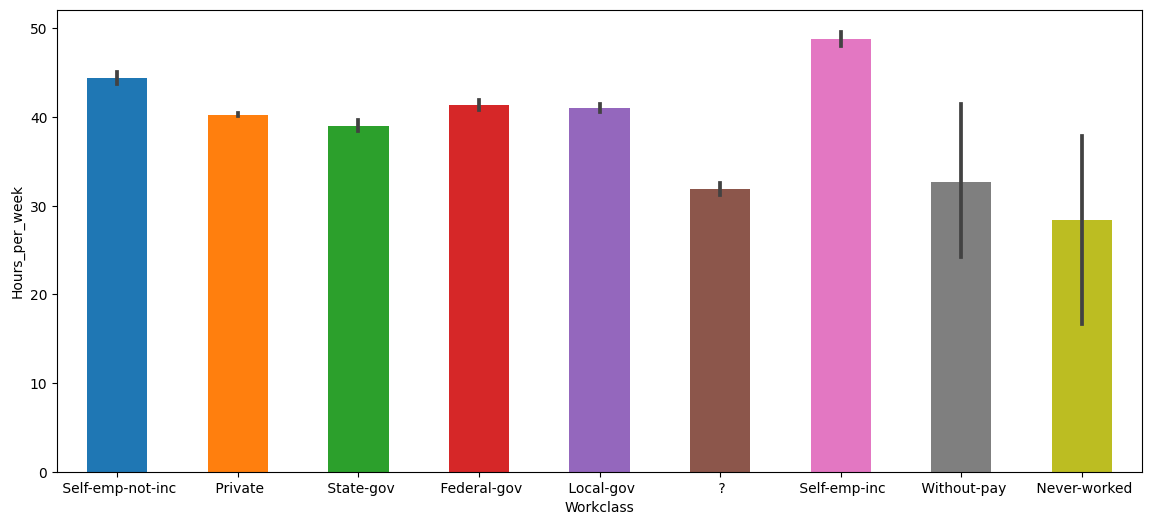
**** ****

**** ****

**Fig:03:data distrbution**

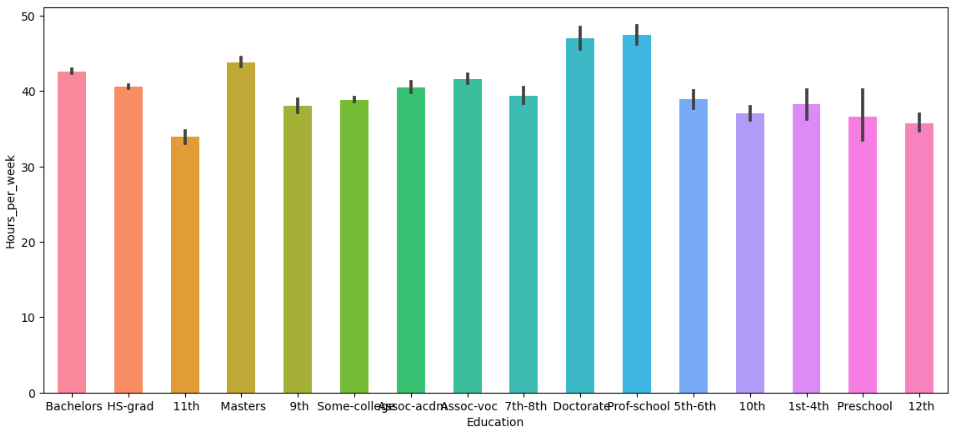
|  |
| --- |
| **Interpretation of Figure :02:**   * **As per graphical analysis we can see that the data distributed of Age column is slightly positively skewed.** * **Fnlwgt is also slightly positively skewed data.** * **Capital loss and capital gain has data positively skewed.** * **Education no and hours per week is almost symmetric data set.** |

**GRAPHICAL PLOTTING OF CATEGORICAL DATA IN RELATION WITH ‘WORKING HOUR PER WEEK’:**

****

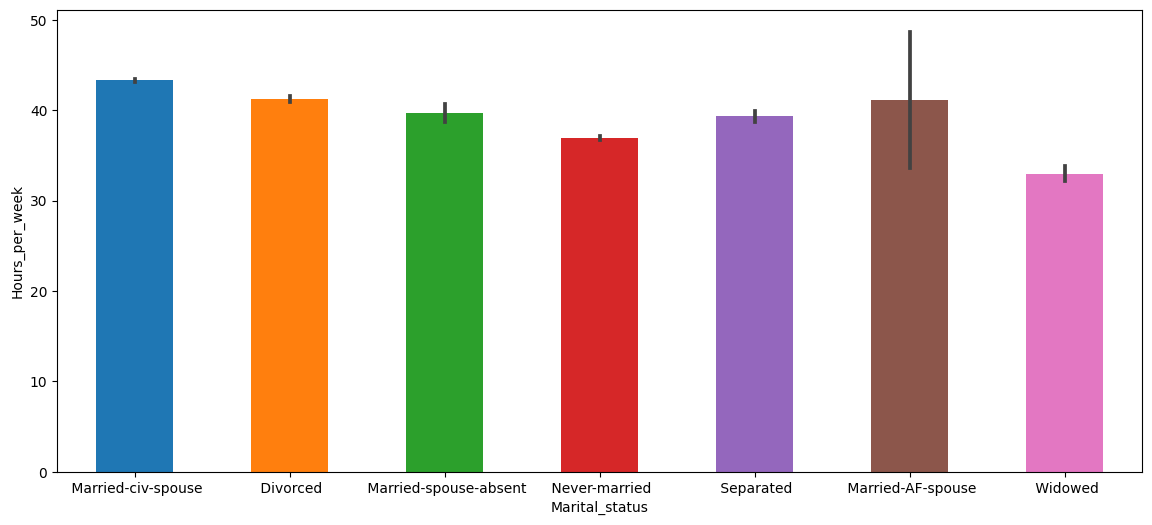
**Fig:04 (Working hour v/s Work Class)**

* As per the above graph we can see that average working hour of self employee and ,local govt work class are higher than other.
* Where as per data set not all self employees are earning more than 50k so there are other categorical factor too involved in earning money.



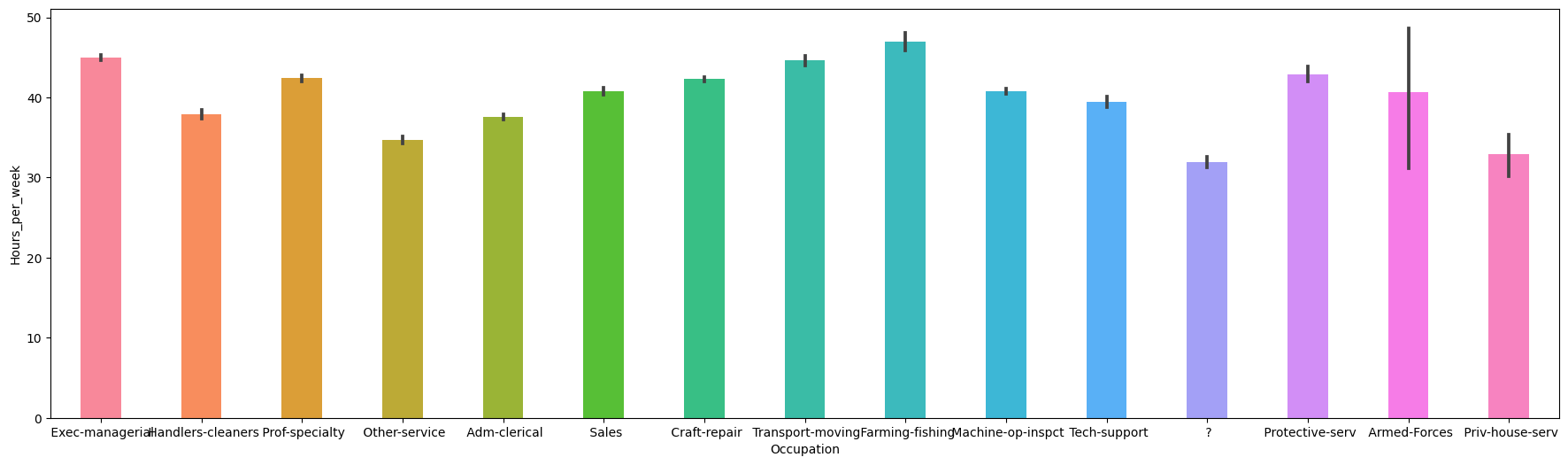
**Fig:05 (Working hour v/s Education)**

* As per the above graph we can see that average working hour of doctor, professor are higher than any other but per data set not all professors and doctors are earning more than 50k so there are other categorical factor too involved in earning money.



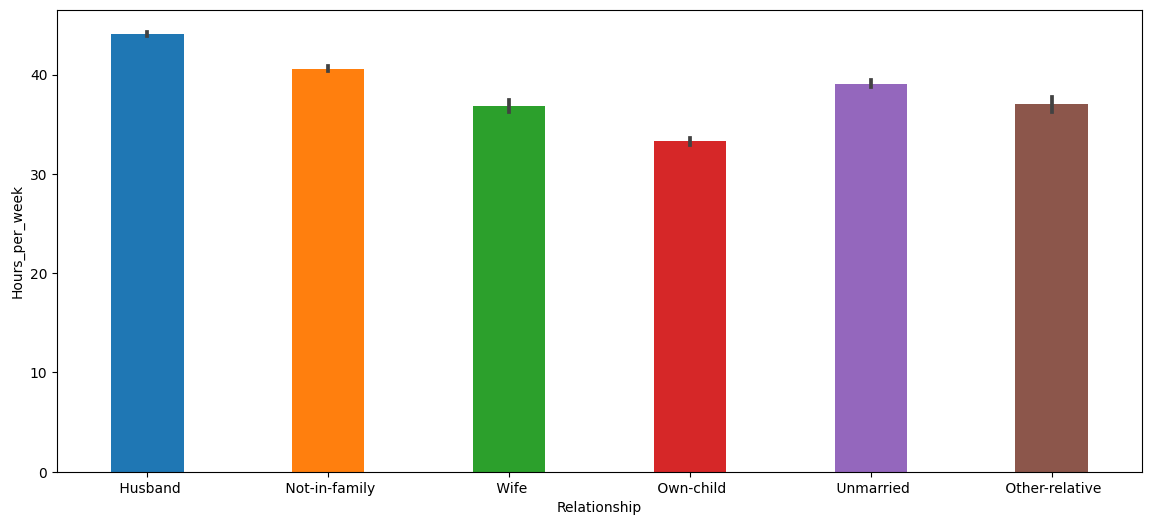
**Fig:06 (Working hour v/s Marital Status)**

* As per the above graph we can see that average working hour of married spouse are higher than any other but per data set not all professors and doctors are earning more than 50k so there are other categorical factor too involved in earning money.



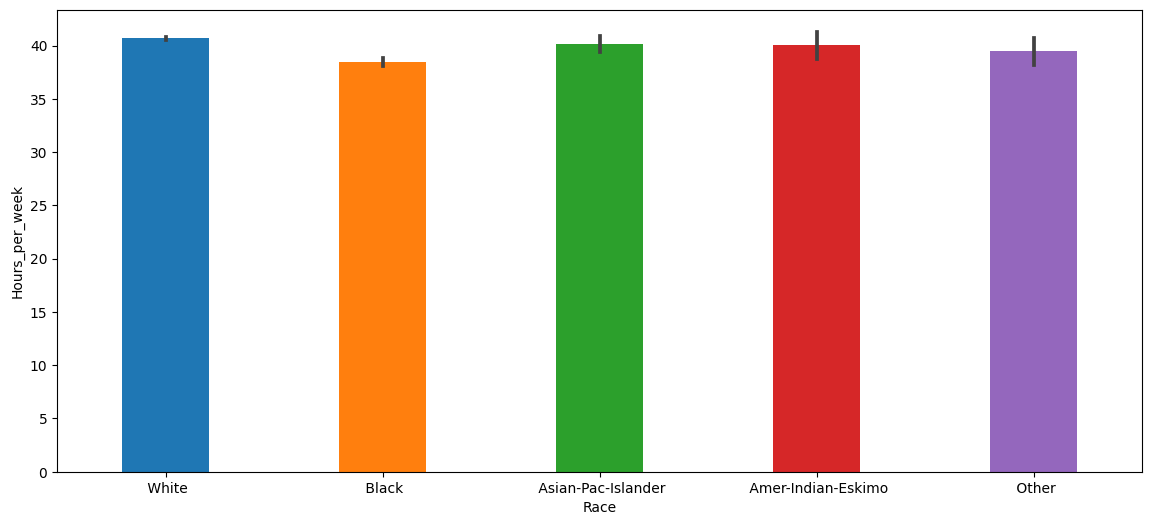
**Fig:07 (Working hour v/s occupation)**

* Farming and fishing is more time taking in terms of working, though it will not give always good return.(over 50K annualy)



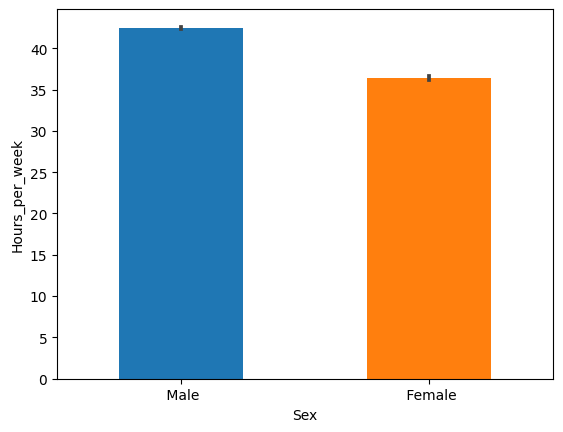
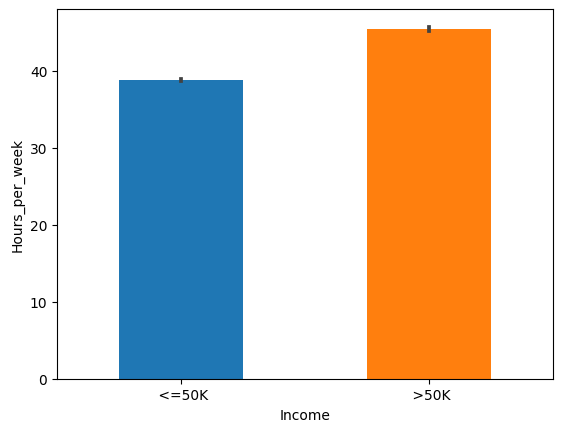
**Fig:08 (Working hour v/s Relationship)**

* As per data , husbands are more investing time to earn money over 50K, but the data set is showing not all husband who working more is earning more.



**Fig:09 (Working hour v/s Race)**

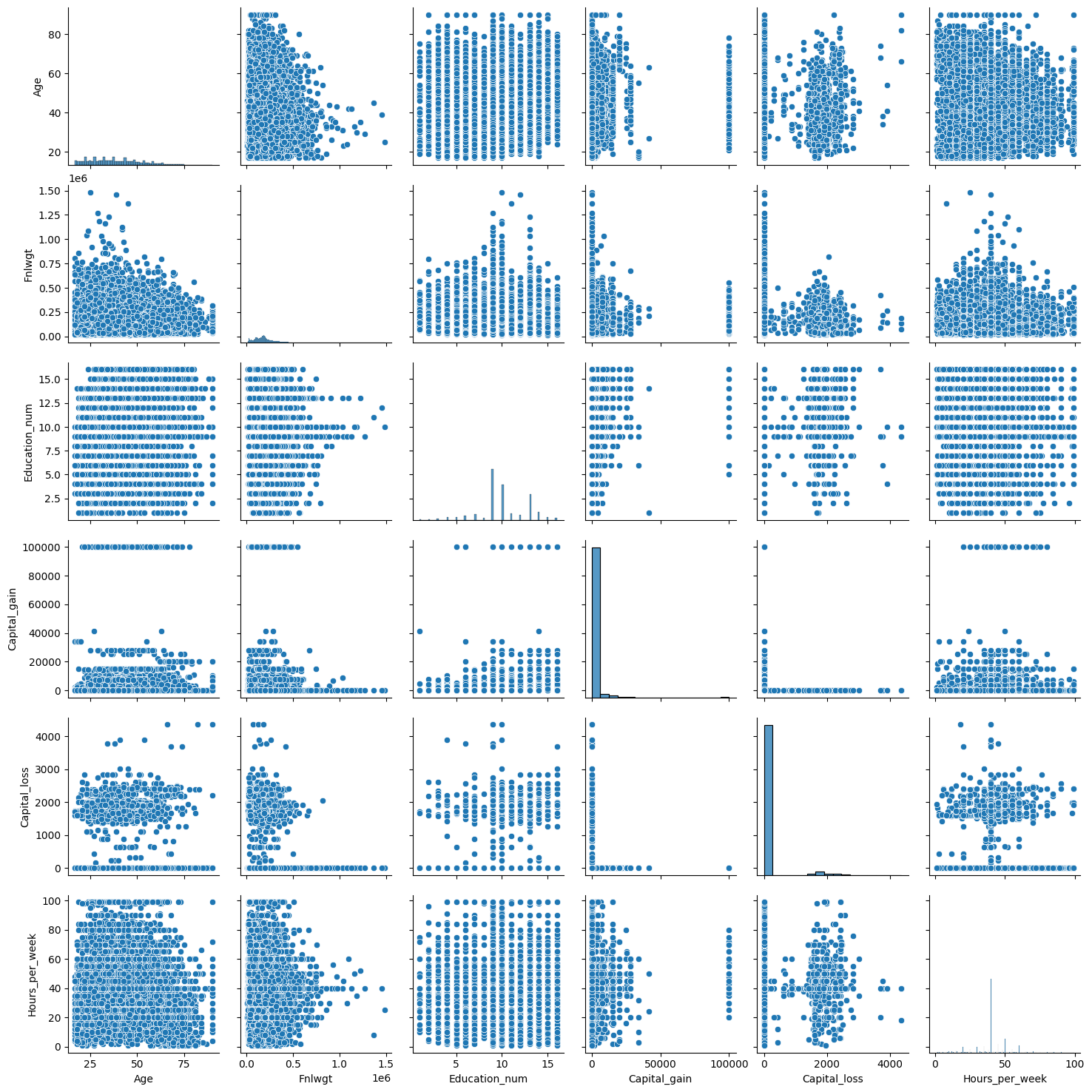
* **Race has almost zero effect on working hour as per graph.**

**** ****

**Fig:10 (Working hour v/s Sex) Fig:11 (Working hour v/s income)**

* Males are harder working than women towards earning money as per plotting.
* Not surprisingly the working hour is directly related with earing money**.**

**PLOTTING PAIR PLOT TO UNDERSTAND THE RELATIONSHIP BETWEEN EACH COLUMN**

****

**Fig:12:Pairplot**

**Interpretation from pairplot:**

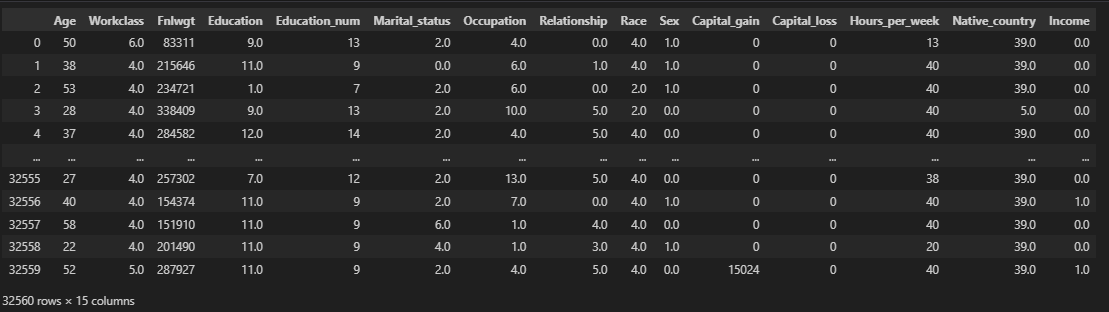
* Age is invert in relation, with Hour per week, fnlwt,as Age increases the these column value decreases and vice versa.
* Working hour is directly related with capital loss and capital gain. It shows positive correlation with each other.
* Education number has positive relationship with capital loss and capital gain column.
* Data is more or less consolidated when we study the relationship of Age column with Capital loss and Capital gain.
* Education number has almost zero correlation with Age.
* Education num has high positive correlation with fnlwt.

With above plot as all the data have their own effect in the data set, we need to consider all the data while doing analysis. We need to be sure after analysis the correlation value.

**DATA MODIFICATION TO FIT THE ALGORITHM**

**Converting Categorical data to numeric data**

**The machine learning is always of numeric value.So we have used ordinalEncoder to convert the categorical data. Now the data set is look like-**

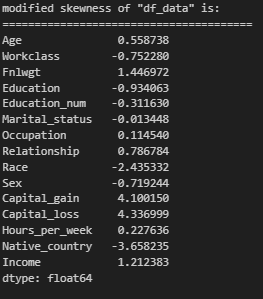
****

**Fig:13:data distribution after application of ordinal encoder**

**Minimize the skewness of the data:**

* **Skewness is related to data distribution.**
* **Outlier is the major issue while analyzing the data. If we do only IQR method there will be a chance of loosing excessive no of data.**
* **Minimizing the skew ness help us to create a level which leads to loss of less amount of data.**
* **Process:**
  + **Analyze the skewness (data.skew() method)**
  + **If the skewness value is more than 05 then we do quberoot of that value to minimize the skewness.**

**Here after minimizing the skewness updated skewness of data set is:**

****

**Fig:14:Skewness**

**OUTLIER DETECTION:**

**For outlier detection we have used Z\_Score method. Z\_score is depending upon the distance of a particular value from mean value. The entire calculation is depending upon the standard deviation and mean value.**

**Formula::**

***Z\_score=(x-mean)/standard deviation…here x is data point***

**To minimize the data loss we have taken the threshold value is ‘5’. If the data point is greater than the threshold value the data point will not be consider for analysis.**

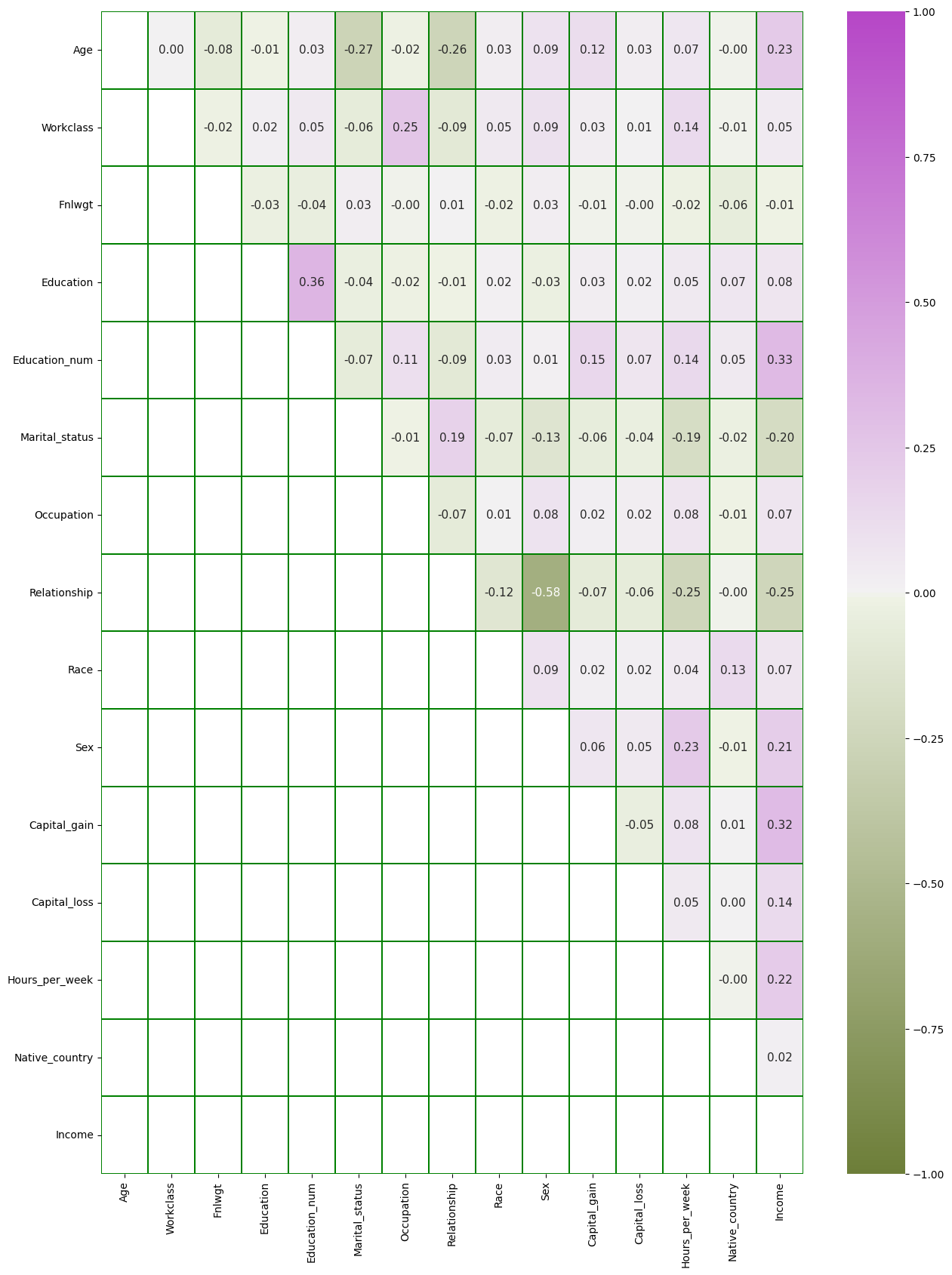
**The data set of Z\_score is like below:**

****

**Fig:14:Z\_score:**

* **As per our criteria for threshold value the data loss of data set will be ‘1.5% approx’**

**Correlation Graph**

****

**Fig :15(Correlation Value of Columns)**

**FEATURE AND LABEL SELECTION:**

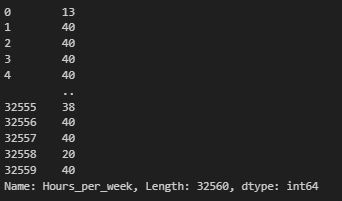
**As we have decided to consider each and every value while analyzing, we haven’t delete any column yet. We are going to create two data set.**

**Data set 01 (x) is for feature and Data set 02 (y) is for label (the value we needs to predict correctly).**

**Data set (x) is look like below:**

****

**Data set (y) is look like below:**

****

**Under standing of influence factor:**

**To get the accurate and unbias result we need to check the varience influence factior. After analyzing we get the below data set of influence factor.**

****

**As per the record we have to remove the highest value of Varience influence factor. So we dropped the column ‘Native\_country’**

**MACHINE LEARNING**

**The data sheet have numeric value with different unit, we have used the Standard Scaler to unified the column value .**

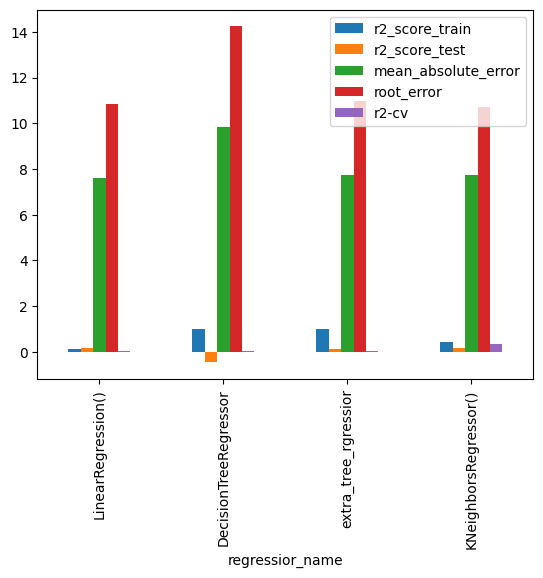
**Splitting data set into Training data and test data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature \_(x) training data** | **Label (y)\_Training** | **Test data (x)** | **Test data (y)** |
| **30% of the data set** | **‘working\_hour’(30%)** | **Rest of the data (70%)** | **‘working\_houre’ (70%)** |

**Random state parameter == 246**

**MACHINE LEARNING MODEL CREATION**

**We have taken Decision tree regressior, Linear Reression technique, Extra tree Regressior technique and Knn technique. Finally after analysis we have got the result. The coparison of the result as below-**

****

**We have chosen the “KNeighbours\_regressior’”provide the highest accuracy**

**Hyper parameter tuning**

* **We have used randomized search CV for hyper parameter tuning and implemented.**
* **After analyzing we get the best parameter of all analysis as below:**

****

* **After all modification we got the best accuracy score ‘17’**

**CONCLUSION:**

**We can conclude that**

* **Option 1: Working hour is not that important for earning over 50k annualy.**
* **Option2: The data set pis not sufficient to analyze the working hour.**
* **Option 3: There is an impurity in data set which crating error to reach accuracy above atleast ‘60’.**