# Demographic and Employment Data for Predictive Modelling

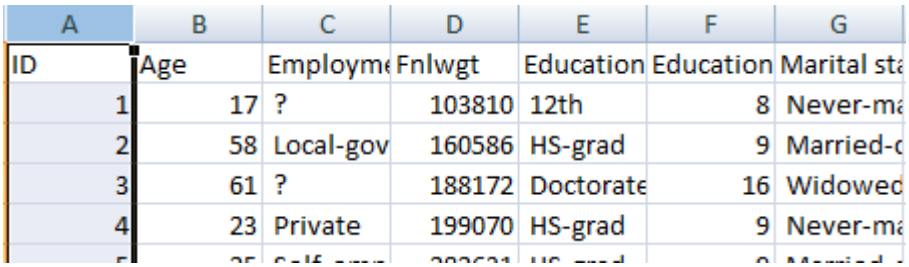
# Contents

# List of Tables

List of Figures

# 0 Preparation of Data

For analysis purpose we have added an extra column of ID. By doing this we would be able to run our data under Knime software.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| | ID | Age | Employme | Fnlwgt | Education | Education | Marital sta |
| | 1 | 17 | ? | 103810 | 12th | 8 | Never-ma |
| | 2 | 58 | Local-gov | 160586 | HS-grad | 9 | Married-c |
| | 3 | 61 | ? | 188172 | Doctorate | 16 | Widowed |
| | 4 | 23 | Private | 199070 | HS-grad | 9 | Never-ma |

Figure 1- Id column is added

# 1A Data Preprocessing

## 1A.1 Identify the type of attributes

The table 1 shows the data attribute characterization based on their values and types

| Attributes | Types |
|---|---|
| **ID** | Interval |
| **Age** | Ratio: Real number quantities |
| **Employment** | Nominal |
| **Fnlwgt** | Ratio: Real Number and represents the number of people in a particular group |
| **Education** | Ordinal |
| **Level** | Interval: Distance is defined and measured in fixed unit |
| **Marital Status** | Nominal |
| **Occupation** | Nominal |
| **Sex** | Nominal: No ordering between them |
| **Capital Gain** | Ratio |
| **Capital Loss** | Ratio |
| **Native Country** | Nominal |
| **Salary** | Ordinal: No distance between values defined and ordering exist |

Table 1- Attributes type Description

## 1A.2 Summarizing Properties

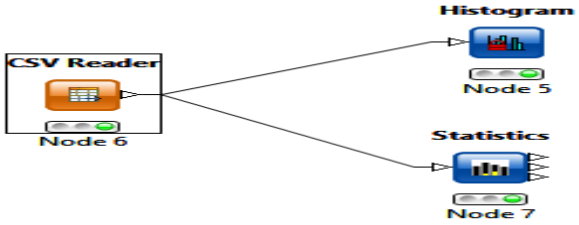



Figure 2- Knime Implementation for Statistics

| Row ID | Column | Min | Max | Mean | Std. deviation | Variance | Skewness | Kurtosis | Overall sum | Median | Row count | Histogram |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | Age | 17 | 90 | 38.444 | 13.6 | 184.97 | 0.574 | -0.224 | 76,850 | 37 | 2000 | 17 90 |
| Fnlwgt | Fnlwgt | 19,700 | 917,220 | 187,271.192 | 102,937.555 | 10,596,140,288.536 | 1.263 | 3.936 | 374,542,383 | 179,579.5 | 2000 | 19.700 917.220 |
| Level | Level | 1 | 16 | 10.072 | 2.505 | 6.273 | -0.387 | 0.796 | 20,144 | 10 | 2000 | 1 16 |
| Captial Gain | Captial Gain | 0 | 99,999 | 1,063.991 | 7,208.102 | 51,956,739.872 | 11.976 | 158.249 | 2,127,982 | 0 | 2000 | 0 99,999 |
| Capital Loss | Capital Loss | 0 | 2,824 | 98.804 | 427.621 | 182,859.953 | 4.235 | 16.537 | 197,609 | 0 | 2000 | 0 2.824 |
| Work Hours | Work Hours | 2 | 99 | 40.342 | 12 | 143.99 | 0.08 | 2.813 | 80,684 | 40 | 2000 | 2 99 |

Figure 3- Values of the summarising Properties

We have used the statistic node to find out the different characteristics of the attributes. The figure 1 shows the implementation on Knime while the figure 2 illustrates the characteristic of each attribute like the minimum and maximum values in each attribute. Also the mean and standard deviation of each attribute can be seen in the figure2.

### 1. Age

| | |
|---|---|
| Min | 17 |
| Max | 90 |
| Mean | 38.444 |
| Std. Dev | 13.597 |
| Variance | 184.877 |

Table 2- Age Attribute Characteristics

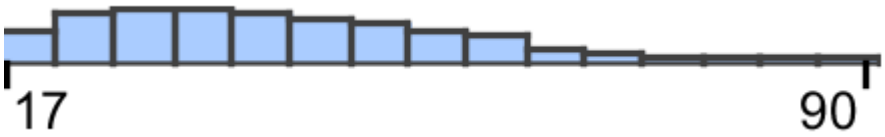



Figure 4- Age Histogram

### 2.Fnlwgt

| | |
|---|---|
| Min | 19700 |
| Max | 917220 |
| Mean | 187271.192 |
| Std. Dev | 102937.555 |
| Variance | 10596140288.536 |

Table 3- Fnlwgt Attribute Characteristics

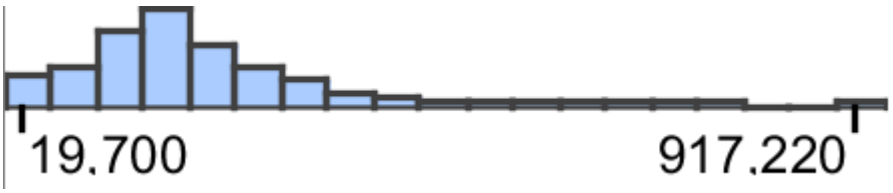



Figure 5- Fnlwgt Histogram

### 3. Education Year

| | |
|---|---|
| Min | 1 |
| Max | 16 |
| Mean | 10.072 |
| Std. Dev | 2.505 |
| Variance | 6.273 |

Table 4-Education Year Attribute Characteristics

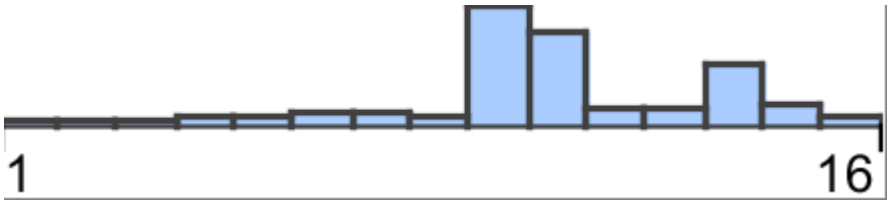



Figure 6- Education Year Histogram

### 4. Capital Gain

| | |
|---|---|
| Min | 0 |
| Max | 99999 |
| Mean | 1063.991 |
| Std. Dev | 1208.102 |
| Variance | 51956739.872 |

Table 5-Capital Gain Attribute Characteristics

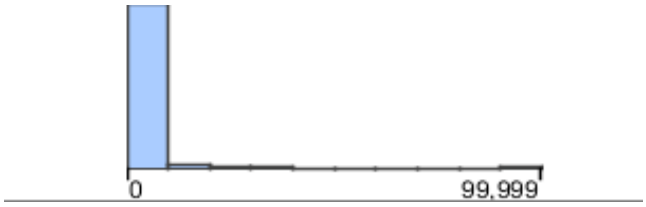



Figure 7- Capital Gain Histogram

## 5. Capital Loss

| | |
|---|---|
| Min | 0 |
| Max | 2824 |
| Mean | 98.804 |
| Std. Dev | 427.621 |
| Variance | 182859.953 |

Table 6-Capital Loss Attribute Characteristics


Figure 8- Capital Loss Histogram

## 6. Work Hours per Week

| | |
|---|---|
| Min | 2 |
| Max | 99 |
| Mean | 40.342 |
| Std. Dev | 12 |
| Variance | 143.99.536 |

Table 7-Hours Per Week Attribute Characteristics

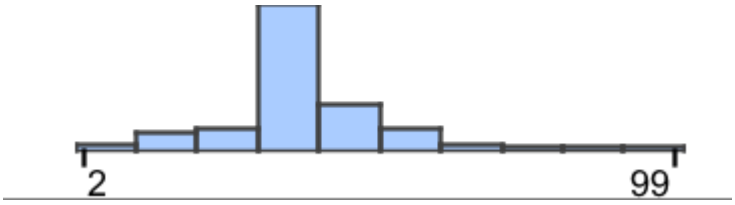



Figure 9- Work hours per week Histogram

The stats shown above, in point 1 to 6 for the different attributes has been processed through Knime Statistics Node. The consequential tables and graphs show the instances

values of specific attribute. By the Histogram graph spread, we can see the minimum and maximum value and their distribution.

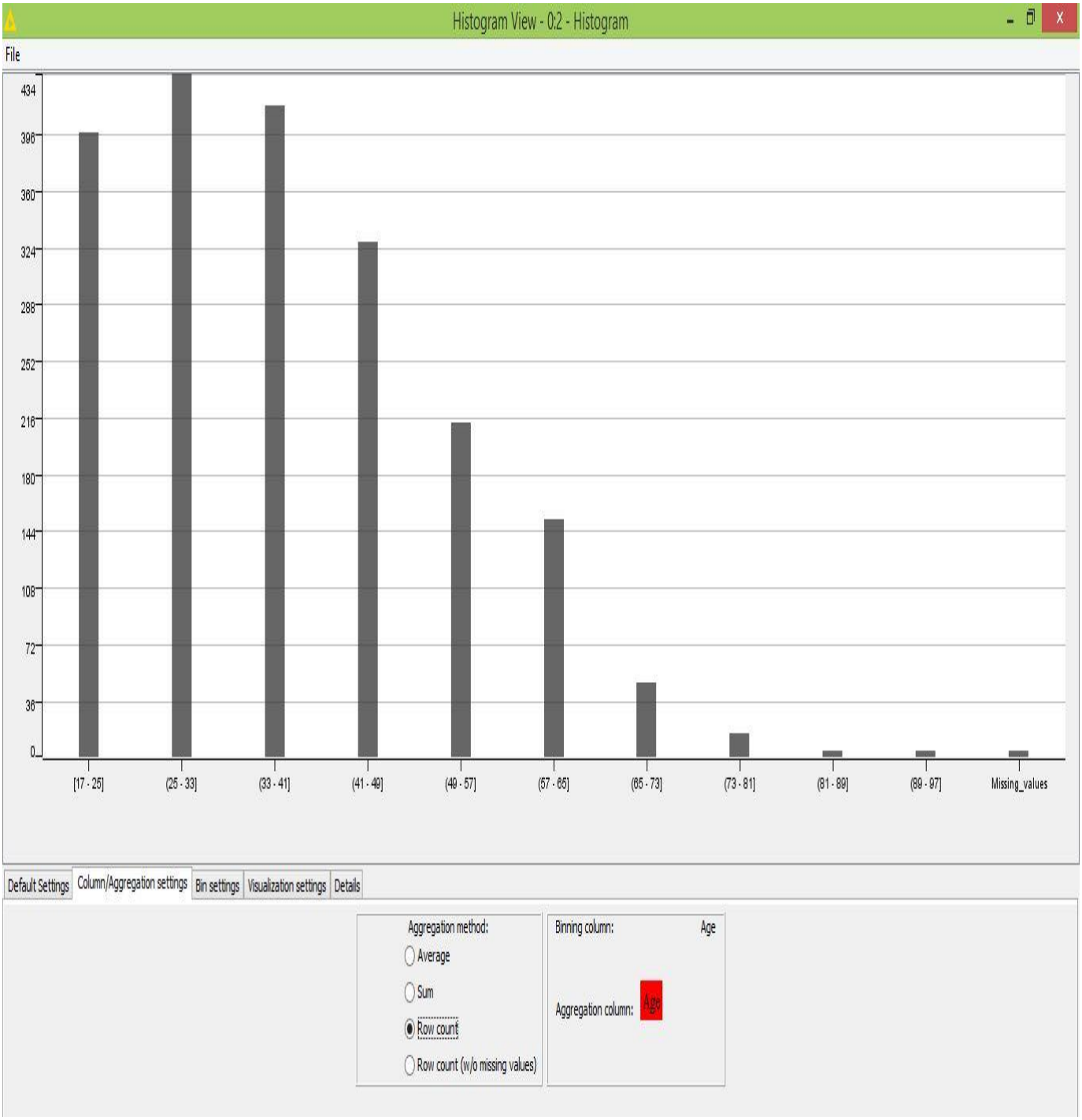



Figure 10-Row Count

Figure 9 show that we can use different aggregate functions to illustrate the data. The row count depicts the number of data in specific attribute. In the above example, we can see the frequency of age like the age between 17 to 25 has frequency of 396.

## 1A.3 Outliers and Clusters

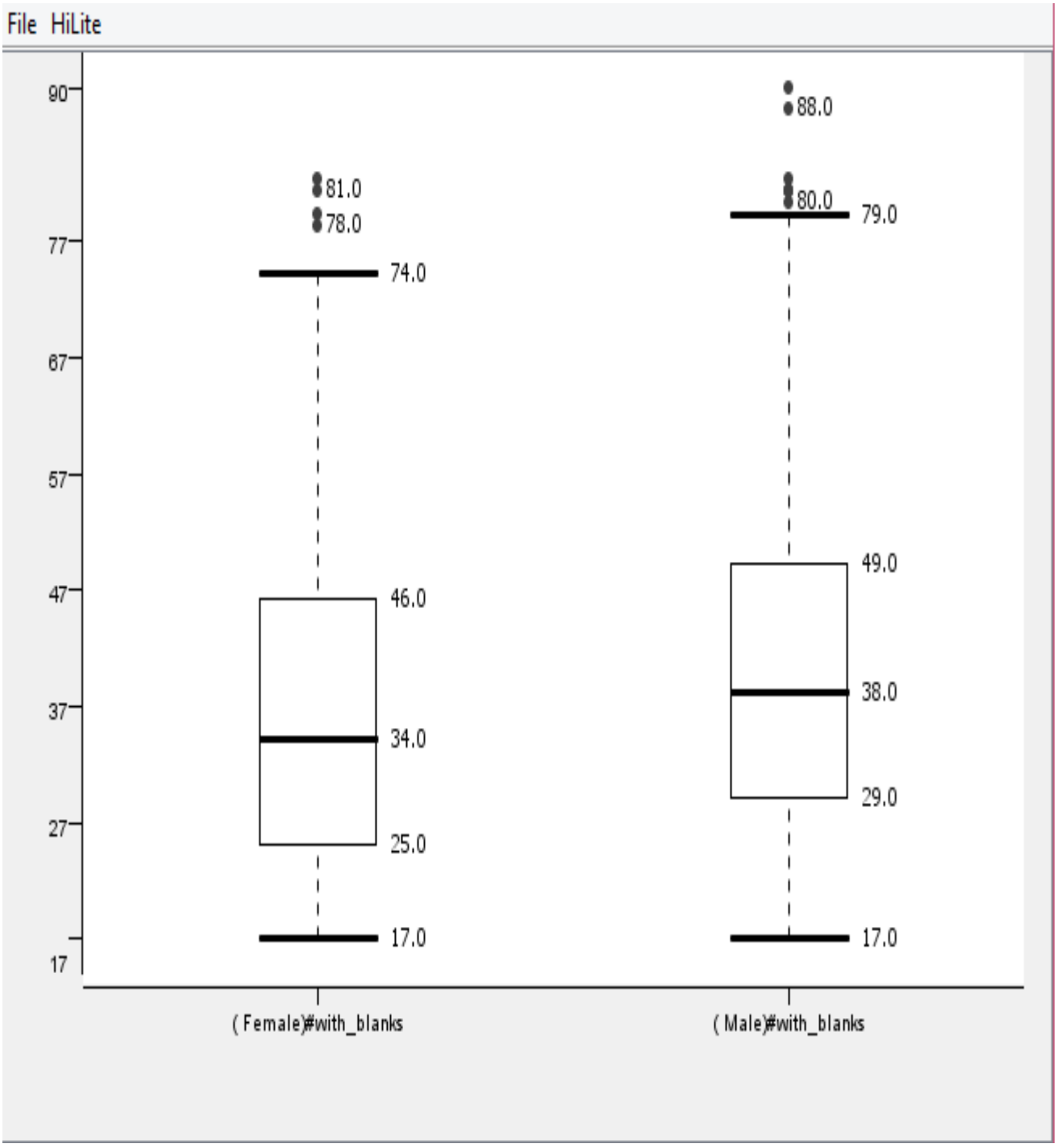



Figure 11-Box Plot Sex According to Age

The Figure 10 show the Box plot that illustrate the starting age of both male and female is 17.But the normal maximum age of female is 74 while male has age of 79.There are some outliers in both female and male that have more age than normal maximum limit.

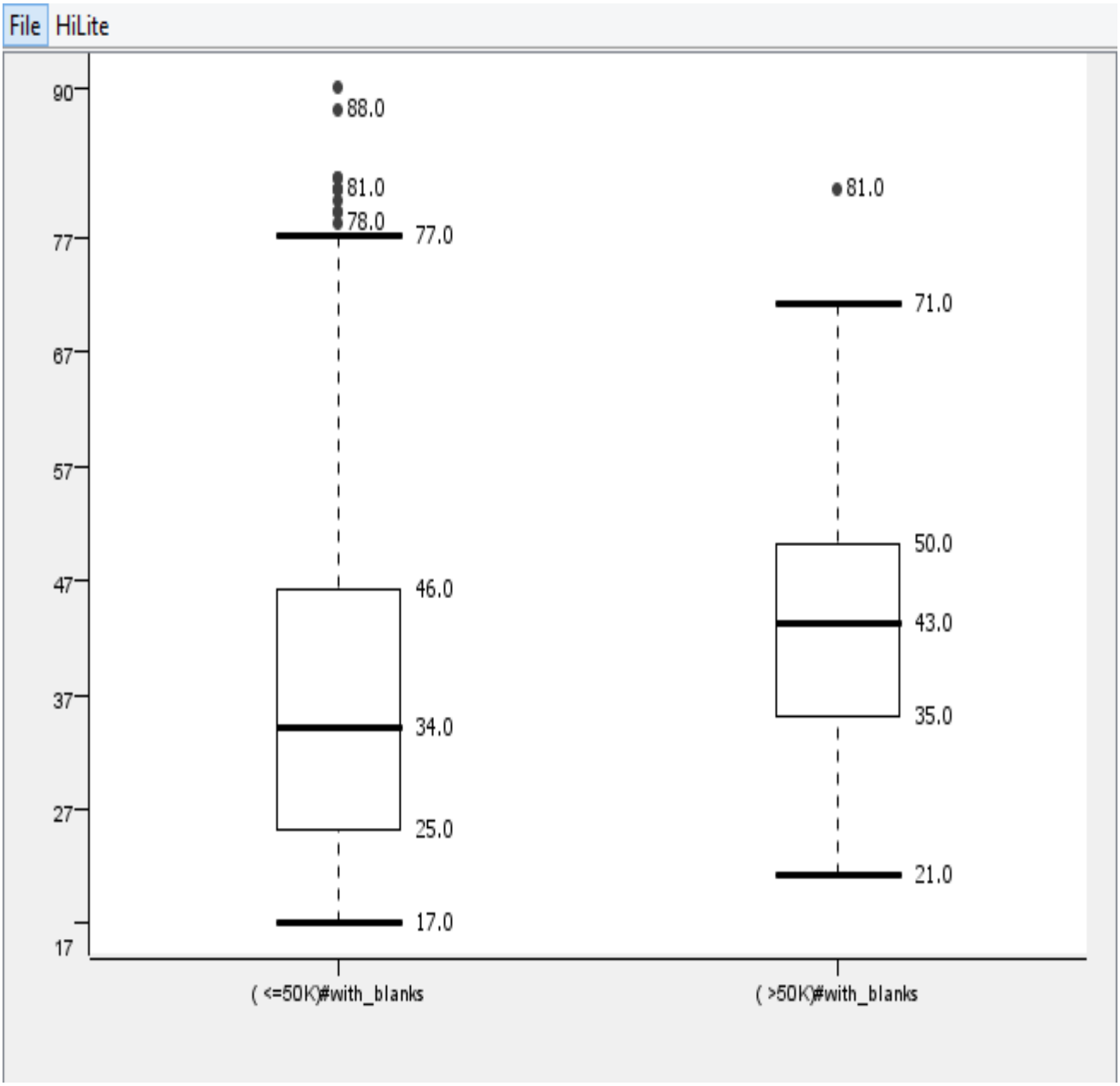



Figure 12- Box Plot of Salary according to Age

In this figure 11, we made the comparison between two attributes one is Salary and the other one is Age. There is only one person who has the salary greater than fifty thousands. Also there are

more people who has age greater than 34 but less than 46 and who has the salary less than fifty thousands.

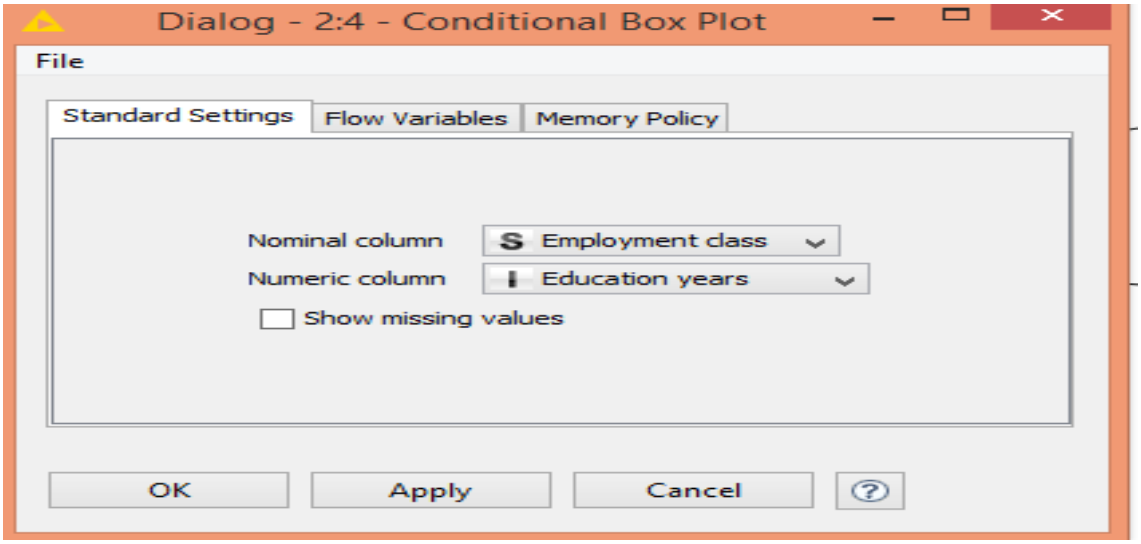



Figure 13- Knime Implementation for Box Plot

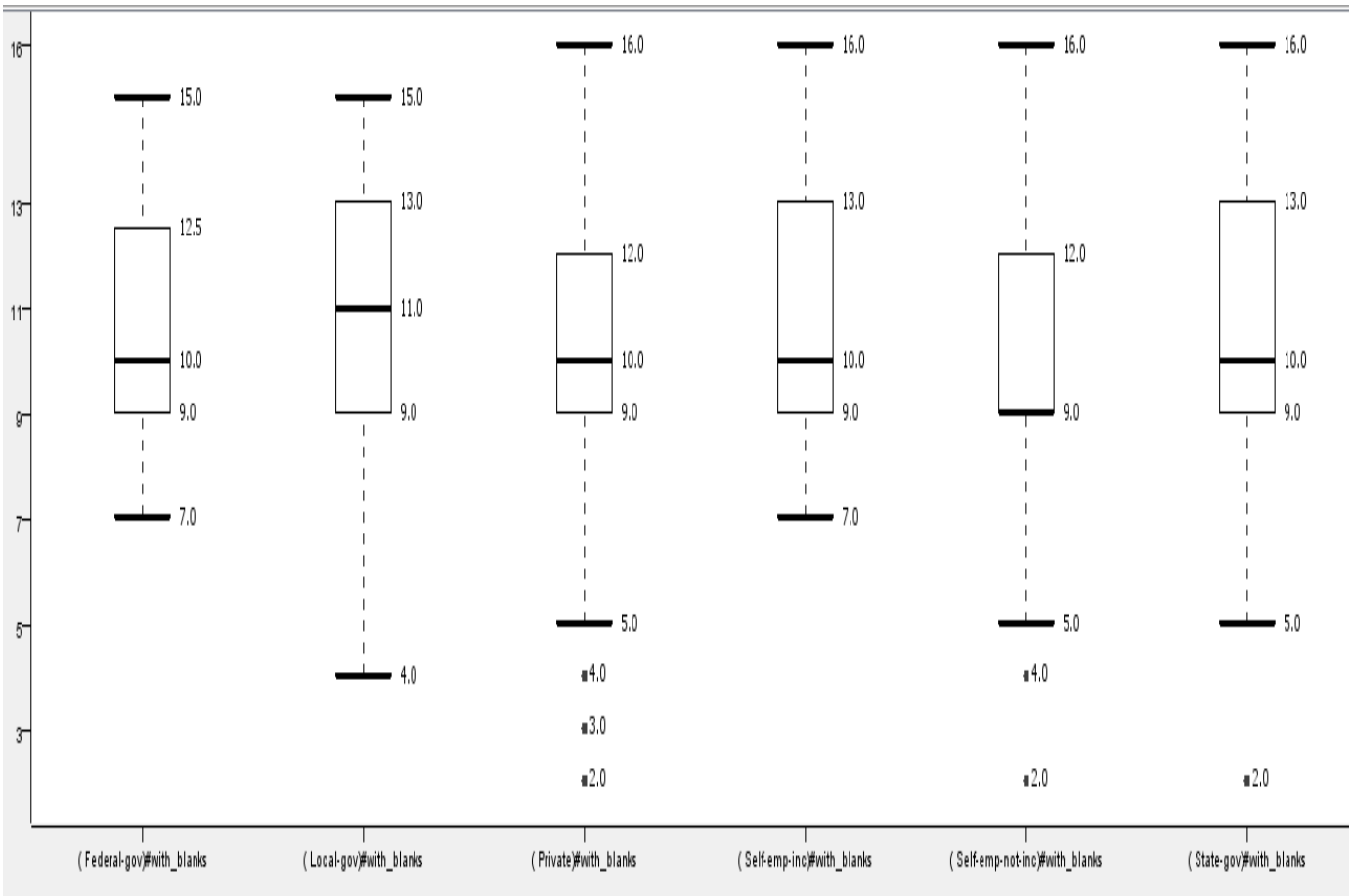



Figure 14-Box Plot of employment according to education years

In figure 13, we drew the box plot of people who are employed a different positions after getting number of years education. For Example more people are self employed who has the ten to thirteen years of education. Also from this figure we can see the education requirement for different employment positions. Like for federal government position the minimum education requirement is 7 years of education while there are more people employed at federal government position who have ten or more years of education. Also in this figure we can see some low outlier at private position.

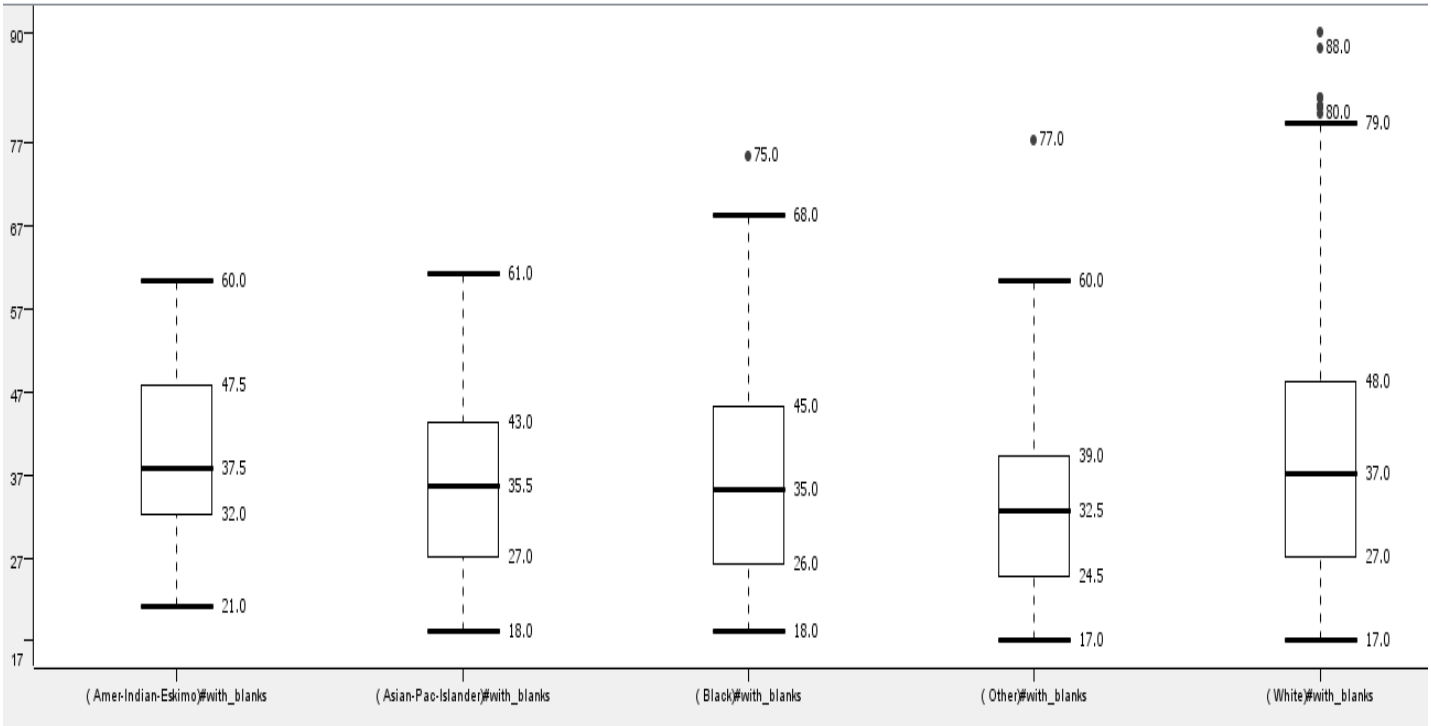



**Figure 15-Box plot of Race according to Age**

In figure 14, the box plot for race according to age is shown that illustrate the average age of white people is 37 years while in case of black people it is 35 years. In case of white people there are some people who have the age over 79 years.

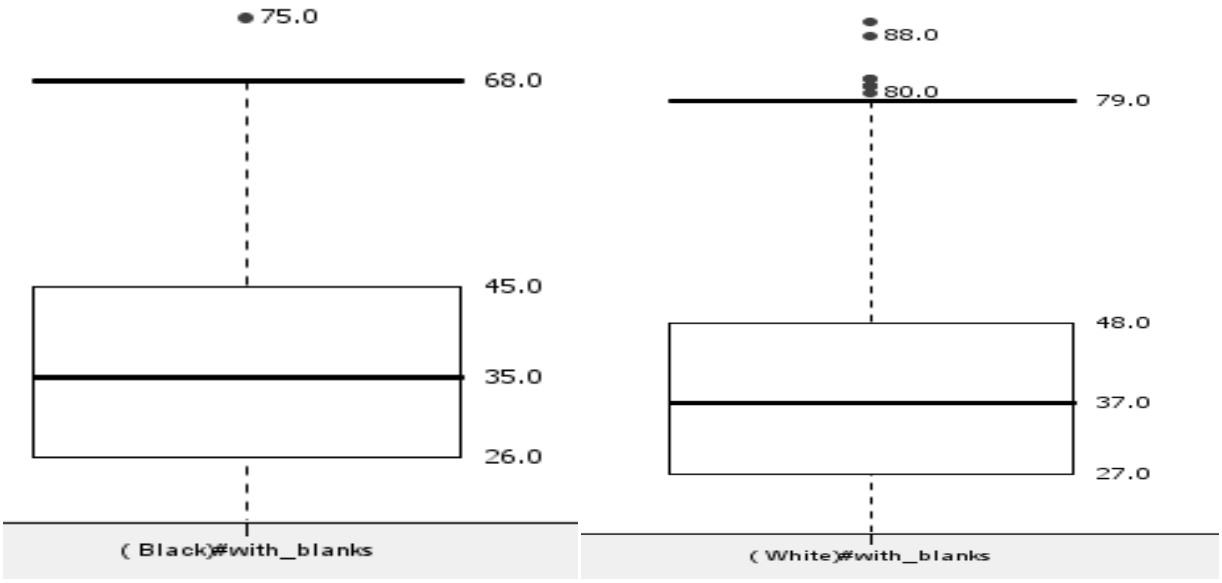



Figure 16-Age comparison according to Race

In Figure 16 a comparison been done between capital loss and gain according to age. From this we observe that people who have age of 65 had more capital gain ratio compared to others.

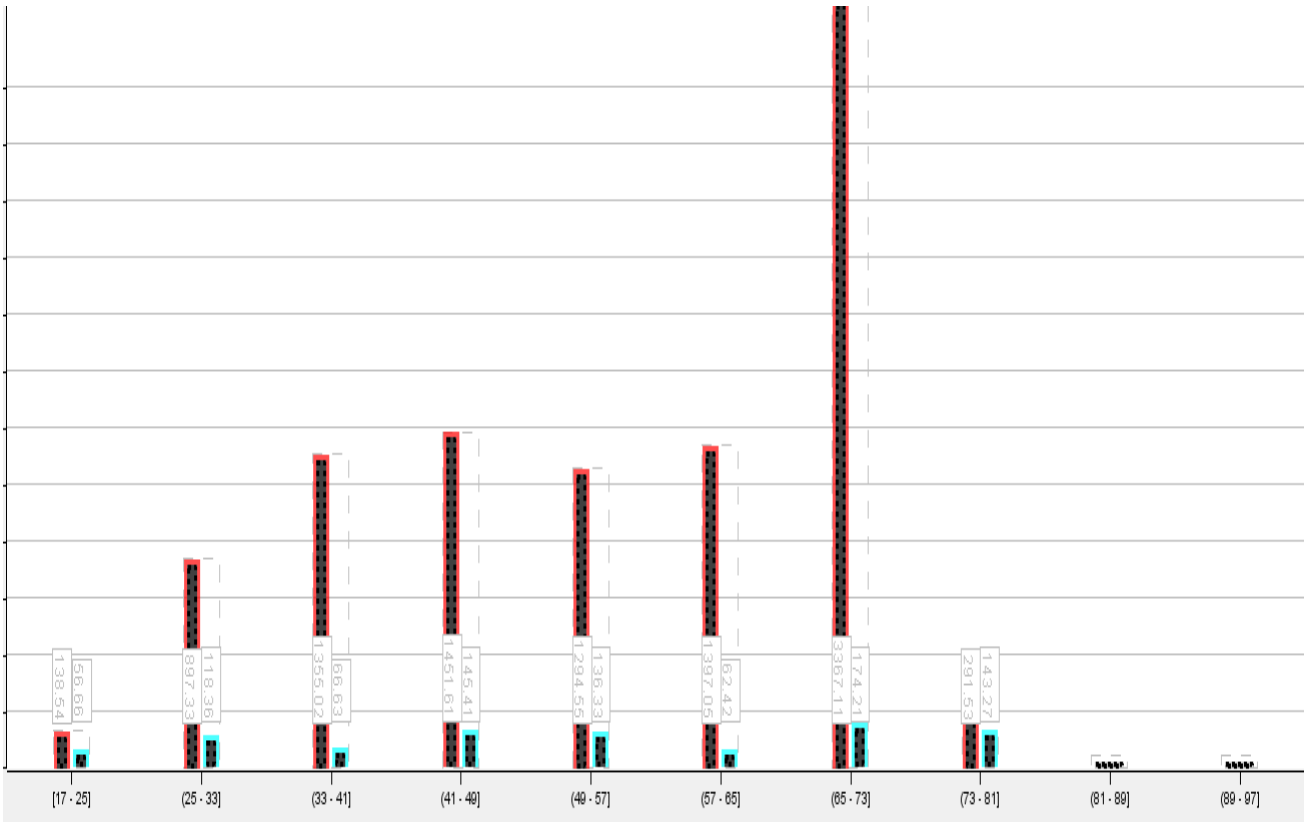



Figure 17-Capital Gain vs Loss according to Age

# Part 1B Data Preprocessing

## 1B.A Binning

We have used equi width and equi depth binning techniques to smooth the data.

| Equi-Width | Frequency |
|---|---|
| **[17,27]** | 508 |
| **[28,38]** | 580 |
| **[39,49]** | 484 |
| **[50,60]** | 295 |
| **[61,71]** | 109 |
| **[72,82]** | 23 |
| **[83,93]** | 2 |

Table 8-Frequency according to Equi Width

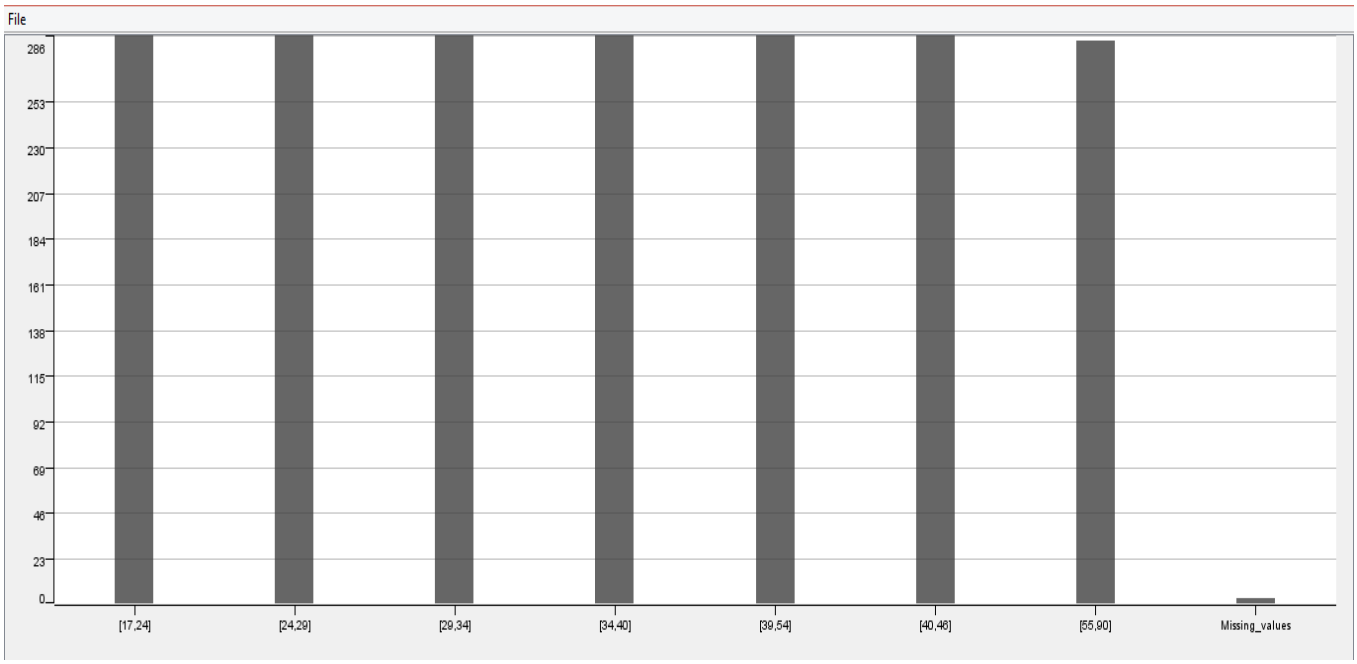



Figure 18-Histogram Equi-Depth according to Age

In figure 11 we can see the euqi depth histogram that shows all data is equally binned.

## 1B.B Normalization

For stable convergences and biases of Age attribute, we have transformed it in same range of values. For this we have used two different normalization techniques

### 1. Min-Max Normalization

In this technique we have used the formula mentioned below

Where A is our data range, Min is the minimum value in our data set which is 17 and maximum value which is 90.Also the newMax value is set to 0 while the newMin value is set to zero.

$$A' = \frac{A - Min}{Max - Min}(newMax - newMin) + newMin$$

Figure 19-Min-Max Normalization Formula

The result for this normalization technique would lie between zero to one as shown in excel sheet. Like for the minimum value in our data set it would be zero and for maximum value it would be one.

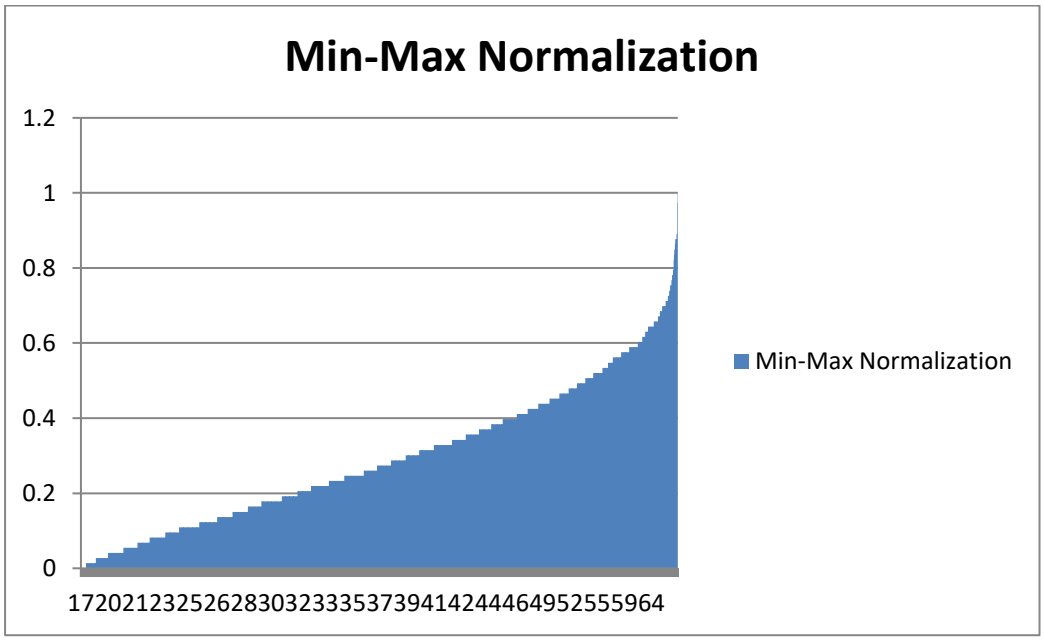



Figure 20-Min-Max Normalization output

Figure 19 shows the output of Min-Max normalization, where y-axis represents the normalized values and x-axis represents the age.

## 2. Z-Score Normalization

$$A' = \frac{A - Mean}{StandardDeviation}$$

Figure 21-Z-Score Normalization

To calculate the Z-Score Normalization we have first calculated the Mean and Standard Deviation and then put them in the above mentioned formula. Where A is our data range, the results can be seen in Excel Spread Sheet.

We can also do the normalization in Knime by using normalizer node as shown in below figure that gives the same results as we have obtained through our manual calculation on excel file

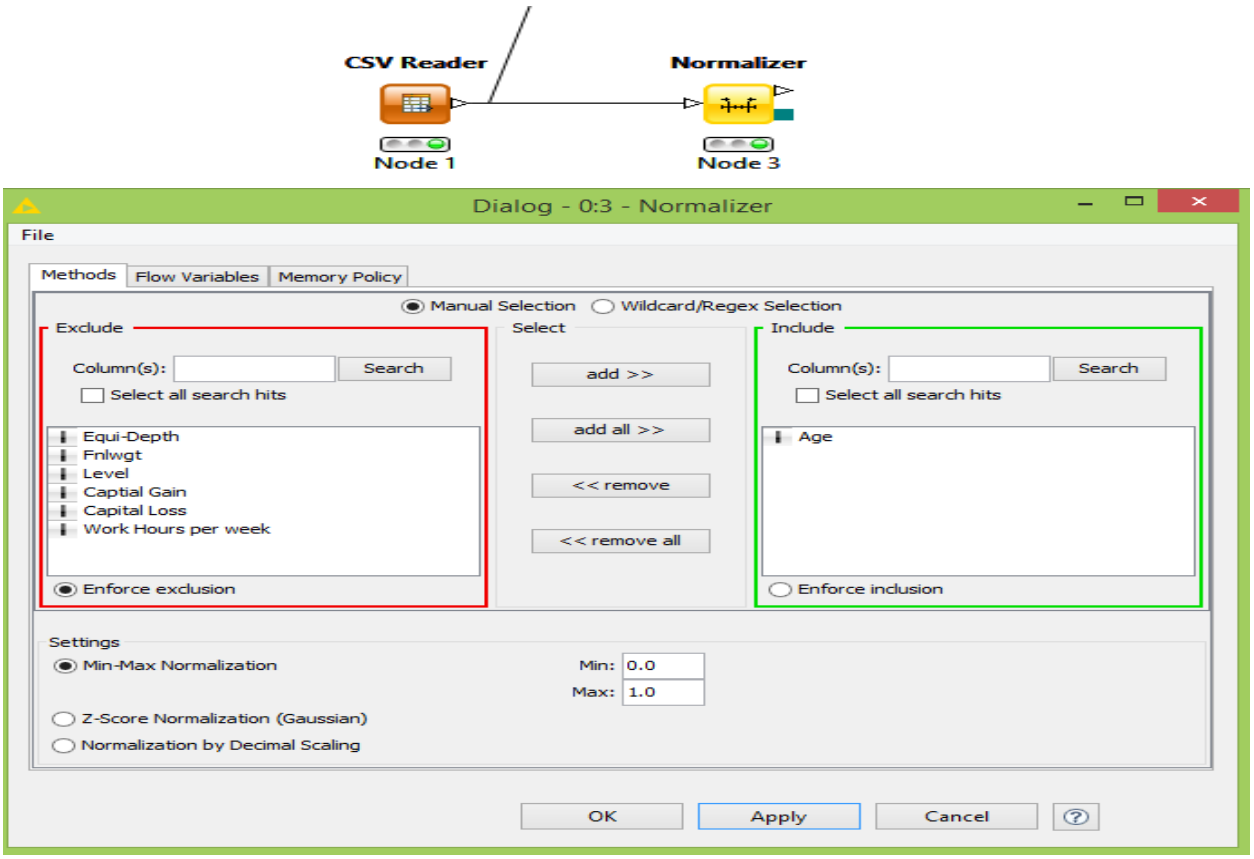



Figure 22- Normalization in Knime

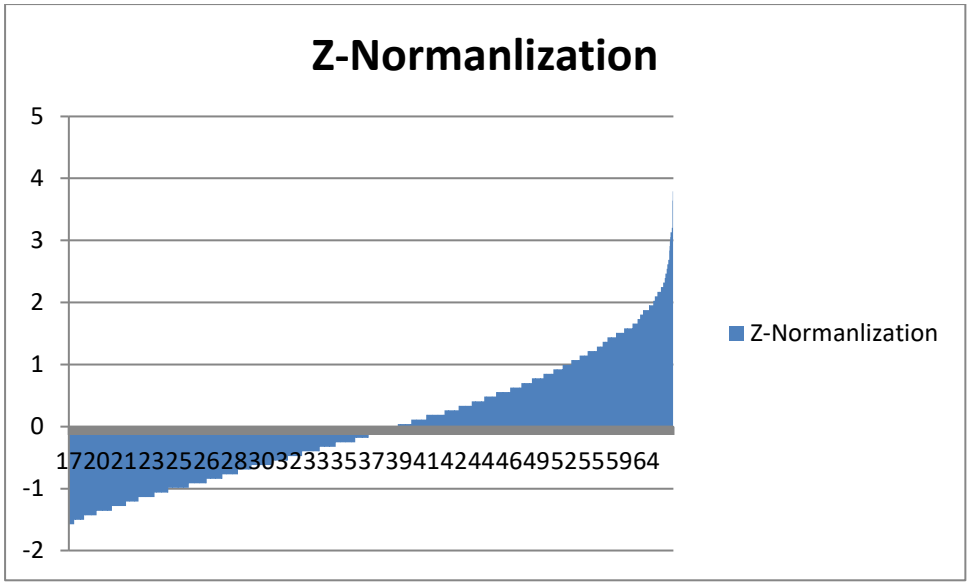



Figure 23-Z-Normalization graphical view

The figure 22 depicts the z-normalized values, where the x-axis represents the values for age and vertical axis represents the values of normalized outputs

## 1B.C Discretize

We have divided the Age attribute in five different categories as mentioned in table

| Category | Range |
|---|---|
| Teenager | 1-20 |
| Young | 21-30 |
| Mid_Age | 31-45 |
| Mature | 46-65 |
| Old | 66+ |

Table 9- Age Category

To achieve this we have used if statement in Excel

IF(A1:A2000<21,"Teenager",IF(A1:A2000<31,"Young",IF(A1:A2000<46”Mid_Age",IF(A1:A2000<66," Mature",IF(A1:A2000>=66,"Old")))))

Also to find out the frequency of each discretise element we have used the countif statement. The results are mentioned in below table and also can be seen in excel file

| Discretise | Frequency |
|---|---|
| TeenAger | 142 |
| Young | 531 |
| Mid_Age | 740 |
| Mature | 520 |
| Old | 66 |

Table 10-Discretise Frequency

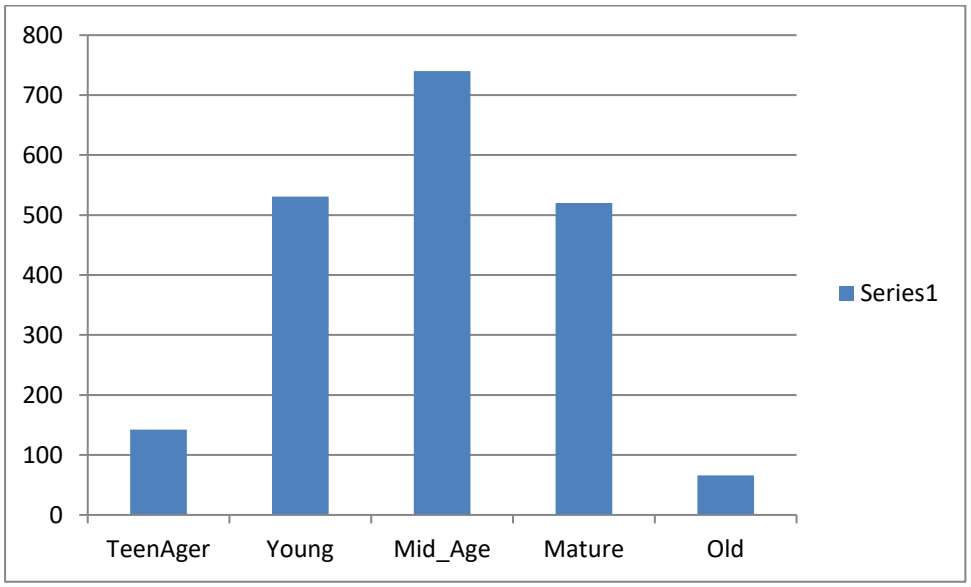



Figure 24-Discretise Frequency graphical view

## 1B.D Binary Conversation

To convert the attribute Level into binary [0 and 1], we first calculated the mean value. And then we have used that mean value as in this case it is 10 as a mid point.

If A < 10 then Value = 0 If A >= 10 then Value = 1

So, the values greater than 10 are set to one while the values below it are set to zero. To achieve this we have applied the following statement in excel

IF(A2:Q2000<=10,"0",IF(A2:Q2000>10,"1"))

Where A represents the Level attribute

## Summary

During our analysis, we found that overall quality of data is very good. There are some attributes like Employment and Occupation who have some missing values in it but it does not affect the overall quality of data. Also there are some attributes like native country and fnlwgt that did not affect the results of classification if been removed from the dataset.

We found the person who have more education or they are more educated earn more. From the data we found that men lives more than female and similarly the white men has more age compared to black men.

During our examination, we scrutinize that certain attributes has quite influenced instances like in case of Capital Loss and Gain they have zero (0) as the leading value and United States as a leading value in Native Country.

From the given data we found that there are more people who have salary less than fifty thousands. Also the people who have more than 10 years of education are either federal employed or self employed.

From our analysis we observed that there is less capital loss at the age of 65 but more capital gain at the same age. This insight can be used to formulate the utmost amount of capital with the probable condition.

By using this data we can predict what sort of people will earn more by using their occupation and demographic details.