CW3 – CST4060

Kinza Shabbir

M00736576

Introduction:

Adult dataset has been selected for this coursework. This dataset contains 32560 observations with 15 attributes. The dataset has been extracted from US census data in 1994. The dataset contains both categorical and integer attributes. The dataset contains missing values[1].

Problem Statement:

1. The purpose of this project is to group individuals based on the demographic information given by adult dataset. This grouping will allow to understand common features within each group.
2. This grouping can be used by marketing companies to target customer within group for a service.

Variable Identification:

The dataset contains the following types of attributes.

Discrete attributes: The discrete attributes only allowing limited number of values to columns. It can be numeric or text.

1. **Workclass:** The work class has information about status of person job.
2. **education:** This attribute gives information about education level of population.
3. **marital_status:** This attribute describe the marital status of a person.
4. **Occupation:** This tells about the information of occupation a person is in.
5. **Relationship:** This describe the relationship of the individual whether married, unmarried or separted.
6. **Race:** The ethnic origin of the individual
7. **native_country:** The individual belongs to which country**.**

Dichotomous : It means the attributes which only have two possible values. The scales of measurement is nominal which means limited number of possible values (Larose, 2005).

8. **Sex:** The gender of the individual**.**
9. **Income:** The information about salary of the individual whether the salary is >50k or <50K

Continuous (Ordinal): Continuous means infinite number of values can be defined within specific range. The following attributes are ordinal because there is order in the values.

10. **age**: The individual age in years

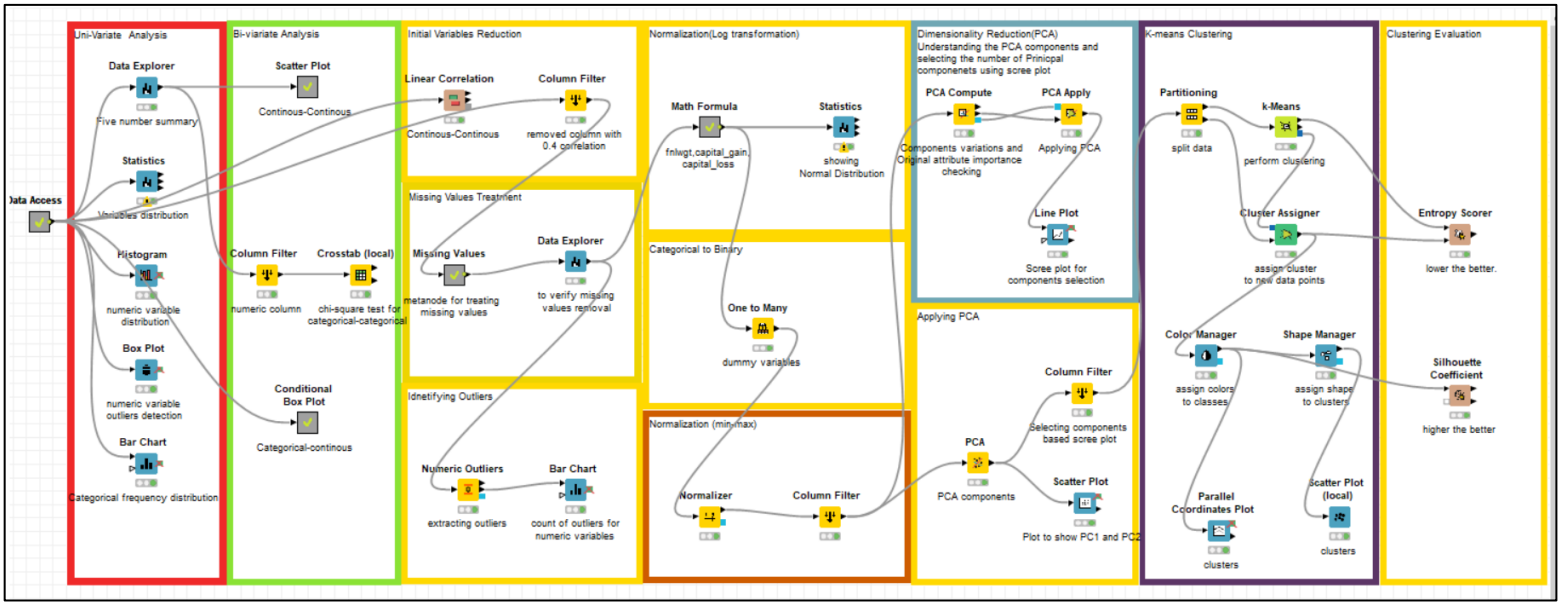
[1] https://archive.ics.uci.edu/ml/datasets/Adult

11. **education_num:** This attribute is giving the same information as education, here it is using numeric value to give ordered information about education level of an individual. So, using only one attribute would be enough for analysis as information is repeated in both columns.
12. **hours_per_week:** Total number of hours an individual works in a week.

Continuous(Numeric):

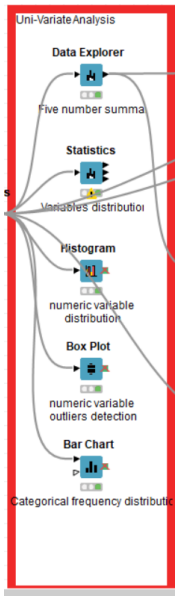
13. **fnlwgt:** This is the final weight assigned after survey based on the type of residence and type of employment an individual have. (Zhu, 2016) The value is greater than 0.
14. **capital_gain:** capital gain score of an individual. The value >= 0.
15. **capital_loss:** capital loss score of an individual. The value >=0

KNIME workflow



Univariate Analysis

Univariate analysis means looking at the individual attribute to understand the distribution of data. The kind of univariate analysis can be applied on attributes depends on whether the attribute is continuous or categorical.



1. **Continuous Variable**

i) Five number summary:

The central tendency and dispersion/variation of data is important in the continuous variables. The summary tables, boxplot and histogram is good choices to do univariate analysis for continuous variables. The following screen shows the summary table for continuous variable and **Data Explorer View** node gives this summary.

- **Min-Max:** The three ordinal attributes **age , education_num and hours_per_week** have limited values as can be seen from the below figure.
- **Variance and Standard Deviation:** tells about the spread of data. **Fnlwgt,capital_gain and capital_loss** have high for variance means the values for these attributes are widely spread. The Standard Deviation is square root of variance.
- **Skewness: fnlwgt, capital_gain, capital_loss** attributes are positively skewed because values are greater than zero. **Age, education_num and hours_per_week** value is almost zero which means that is normally distributed but not completely.
- **Kurtosis: age, education_num** has flat peak as these attributes have value close to zero for Kurtosis whereas **fnlwgt, capital_loss,capital_loss, hours_per_week** have higher value than zero which indicates these attributes have high peak in the distribution**.**
- **No. missing:** There is no missing or No NaN for these six attributes.
- **No. Zeros:** capital _gain and capital loss have so many zero values in the attributes.

Data Explorer View

Numeric | Nominal | Data Preview

Search:

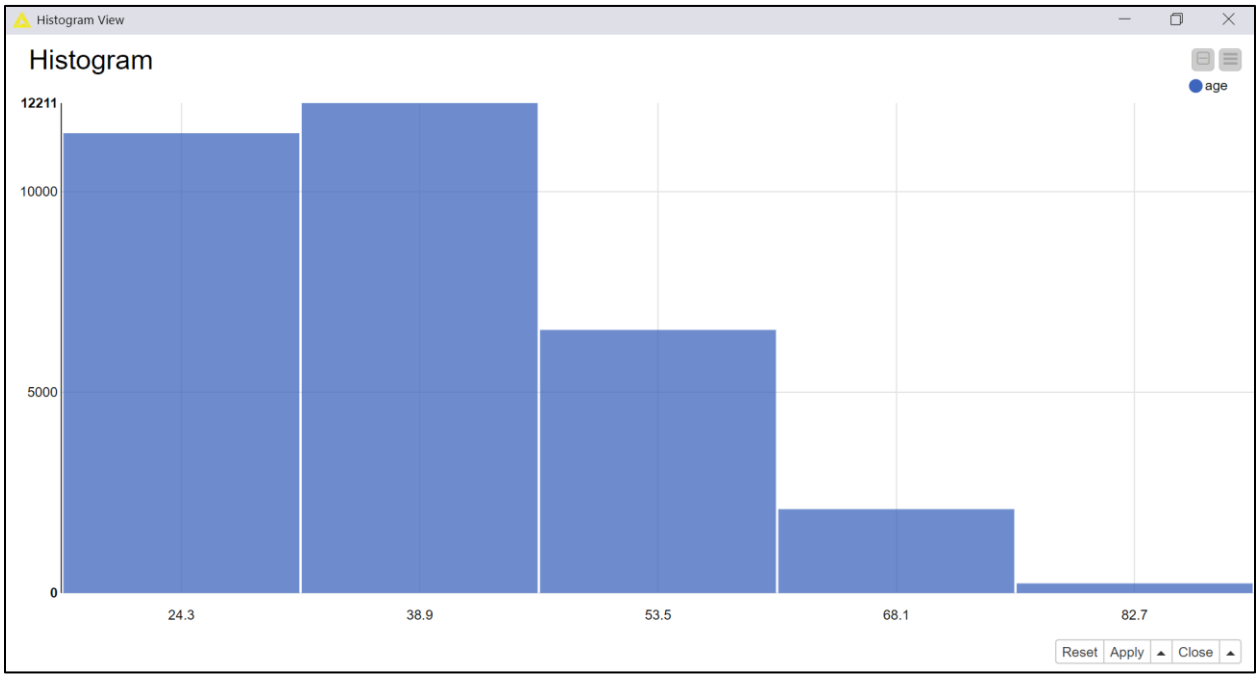
Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN
age		17	90	38.582	13.640	186.061	0.559	-0.166	1256257	0	0	0
fnlwgt		12285	1484705	189778.367	105549.978	11140797791.842	1.447	6.219	6179373392	0	0	0
education_num		1	16	10.081	2.573	6.619	-0.312	0.623	328237	0	0	0
capital_gain		0	99999	1077.649	7385.292	54542539.178	11.954	154.799	35089324	29849	0	0
capital_loss		0	4356	87.304	402.960	162376.938	4.595	20.377	2842700	31042	0	0
hours_per_week		1	99	40.437	12.347	152.459	0.228	2.917	1316684	0	0	0

Showing 1 to 6 of 6 entries

ii) **Histogram (Frequency Distribution)**

The **histogram Node** has been used to show frequency distribution among different ranges.

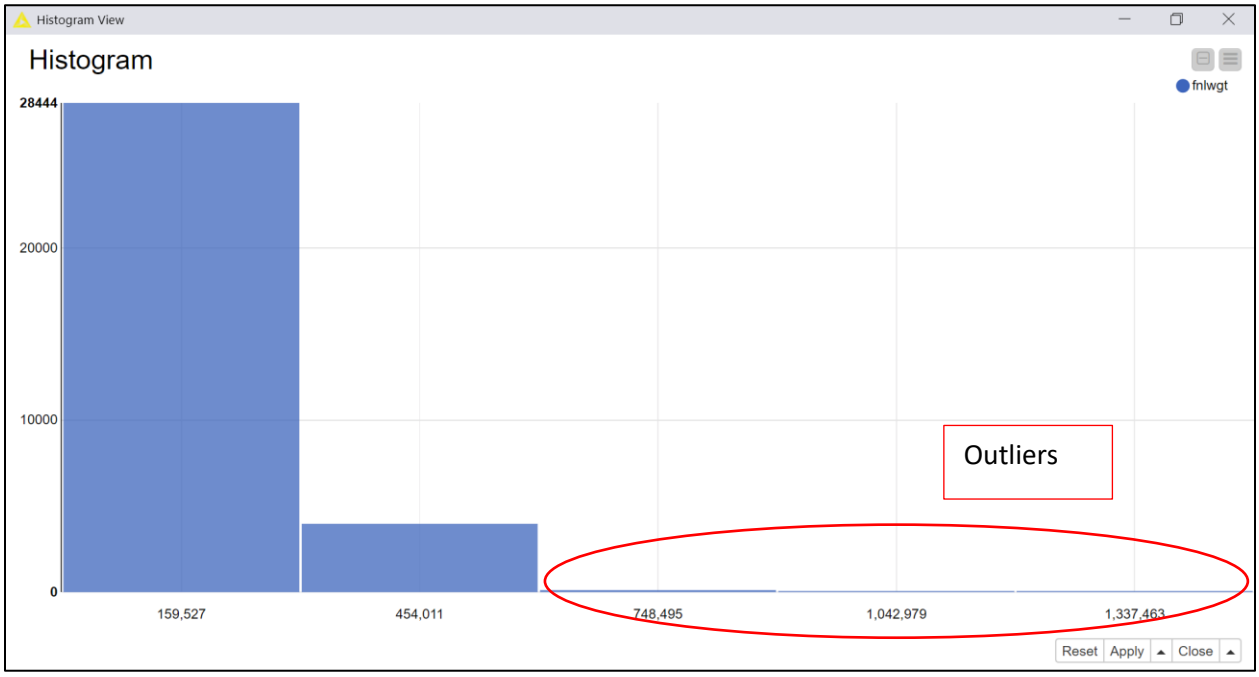
- **age:** The histogram below shows the highest number of values of age centred around 38.9 and frequency=12211 whereas lowest frequency=241 are centred around age 82.7. The histogram doesn't have bell shape distribution but still the data is approximately distributed with slightly positively skewed.



Fnlwgt:

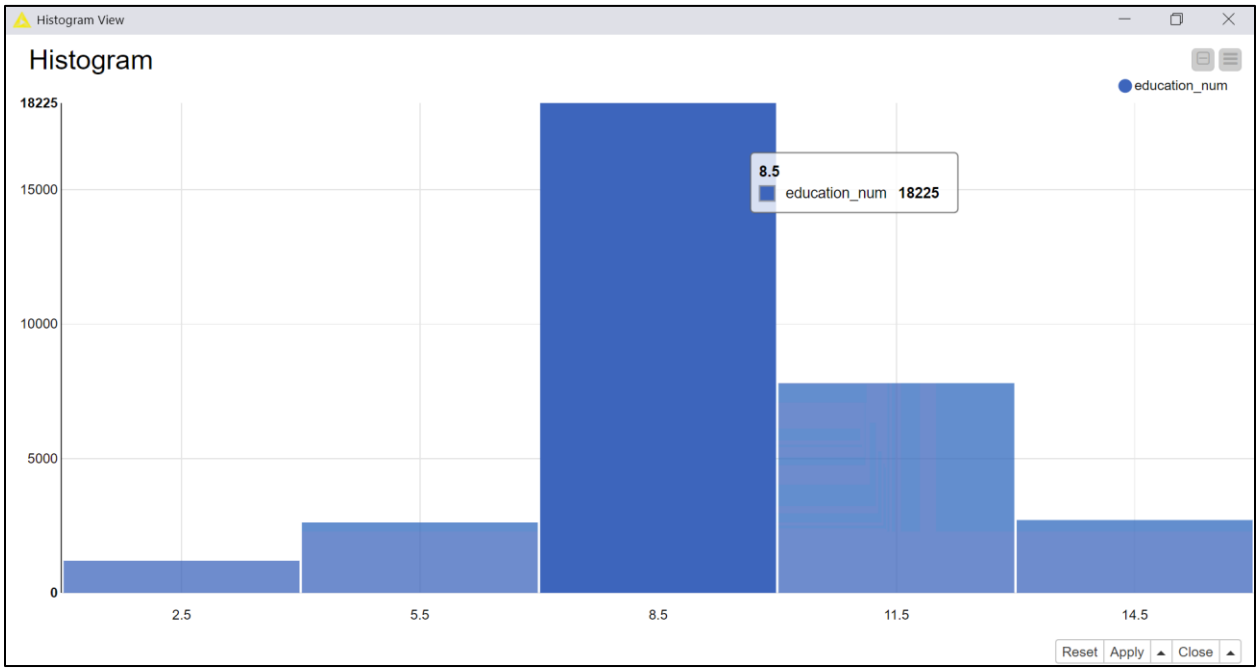
- The data is positively skewed and have very high peak
- The highest values=28444 are centred around 159,572.

- The histogram also shows outliers in last distribution which has 15, 5 values in the last two distribution respectively.



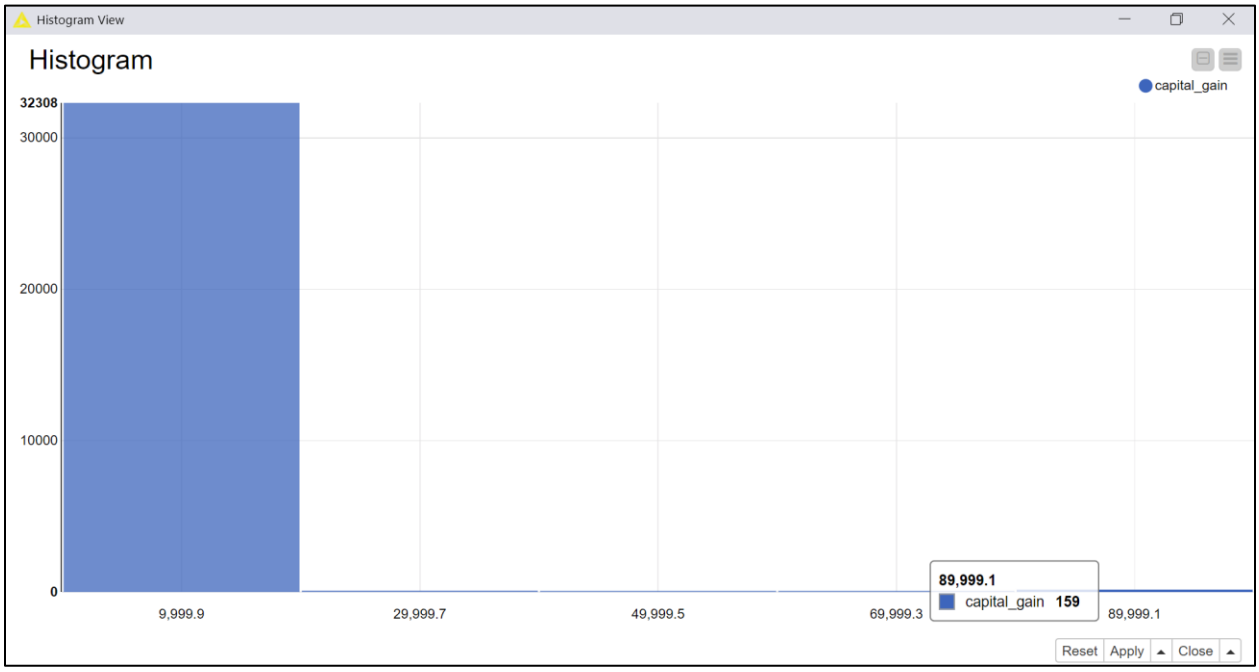
Education_num

- The education_num attribute is almost equally distributed as can be seen from histogram with highest values in the centre which 18225.
- There is no skewness and outliers.



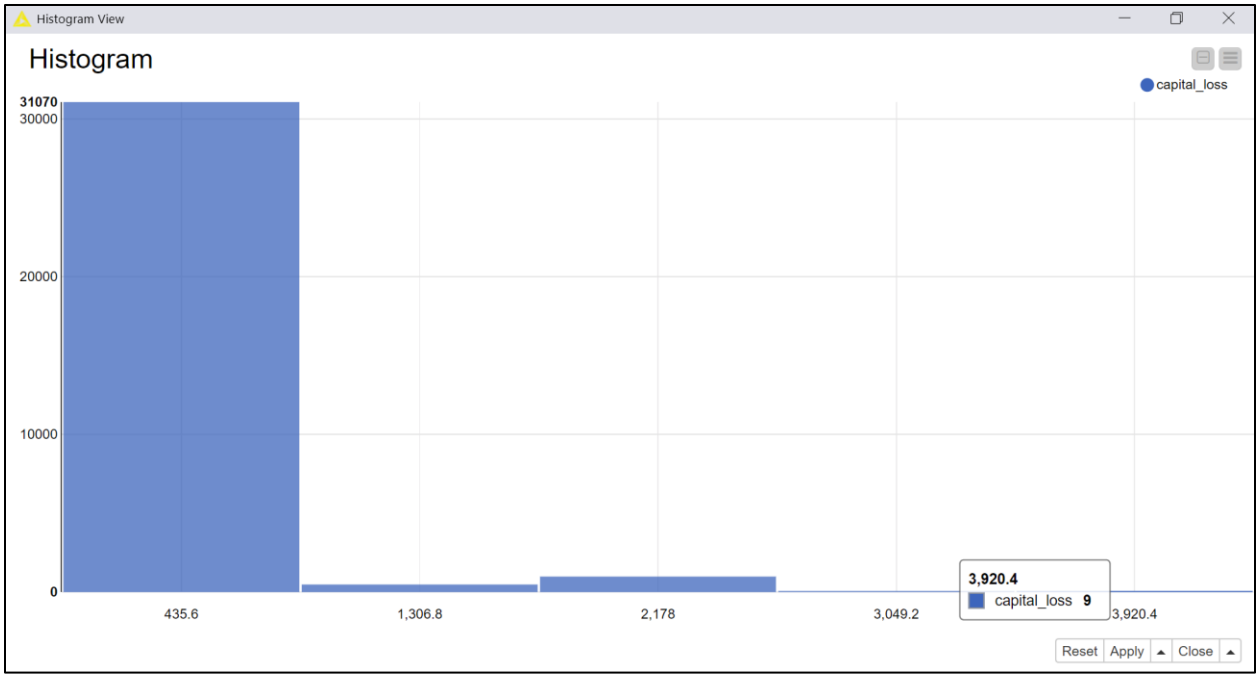
Capital_gain:

- The data is positively skewed which can be seen from the histogram.
- 32308 values are distributed around 9,999.9 value.
- There are 159 outliers whose values are almost 9 times greater than the average value.



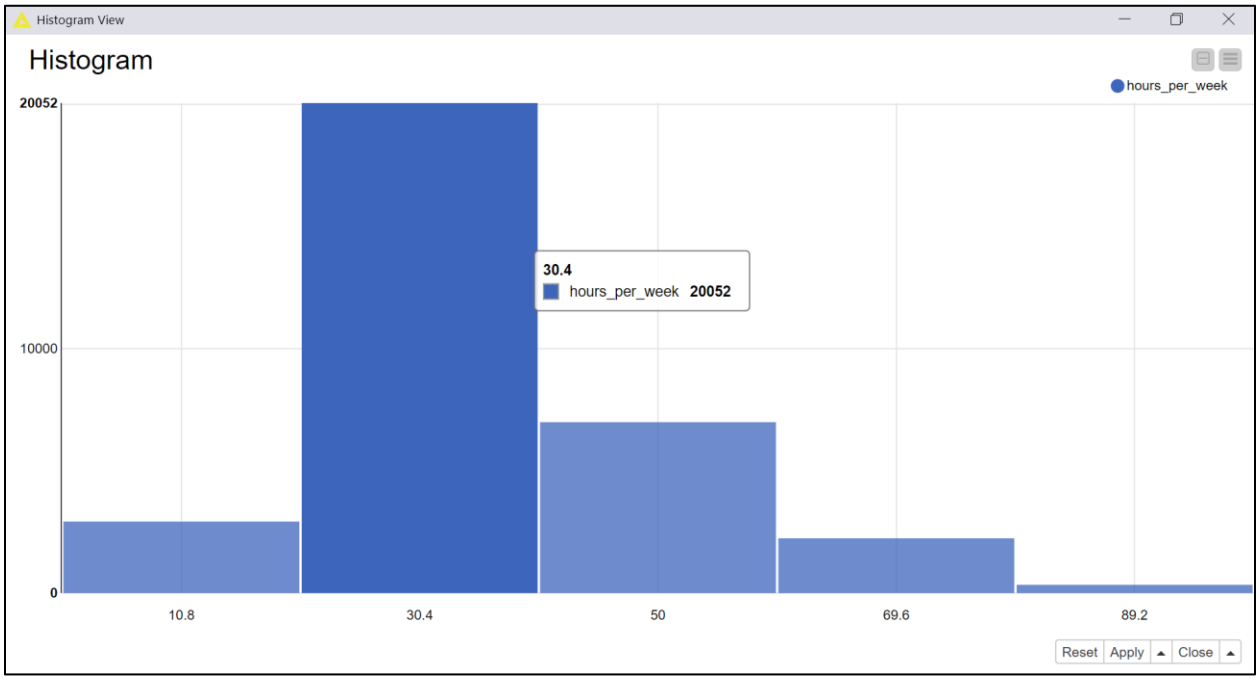
Capital_loss:

- **Capital_loss** attribute is positively skewed. There is need to apply transformation to make it normally distributed.
- **31070** frequency of values lie around **435.6**.
- **9** values are eight times higher than average value which are considered as outliers.



Hours_per_week

- Approximatly normally distributed.
- 20052 frequency of values lie around 30.4 hours per week.
- There is no outliers and no need to apply transformation for distribution.



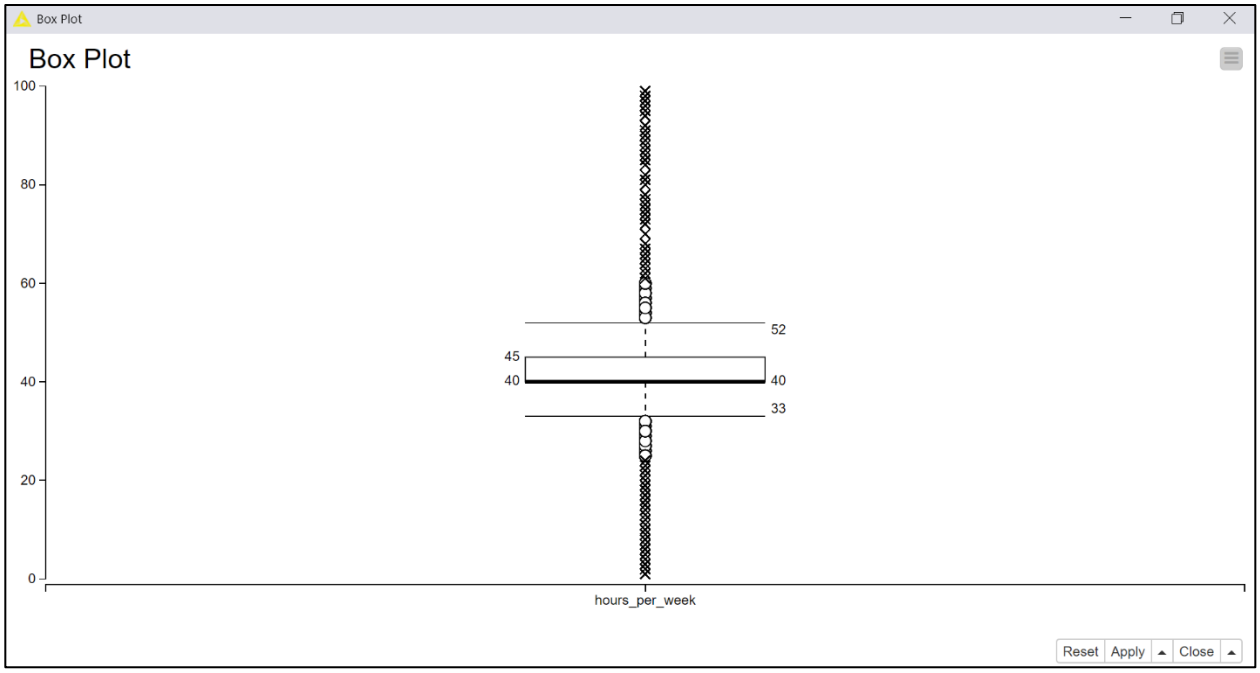
iii) Boxplot (Outlier detection)

hours_per_week:

This box plot is good to visualize the outliers exists in the attributes.

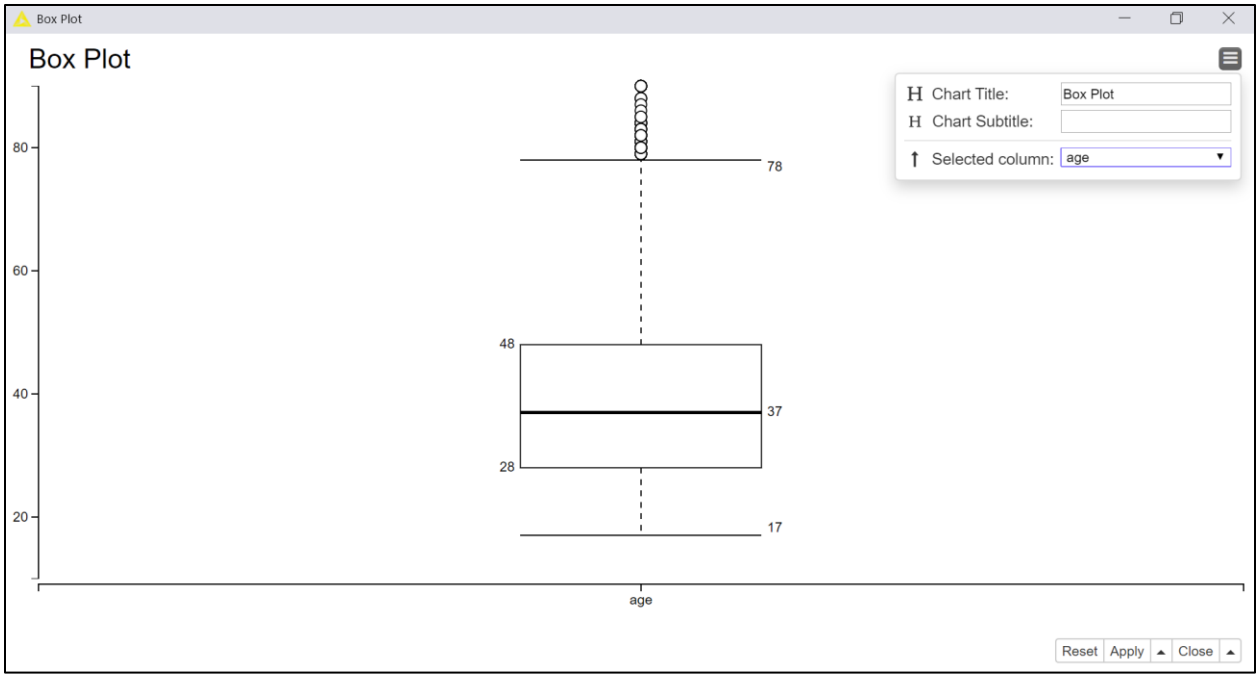
Outliers detection: **Q1-1.5*IQR= 4.5(lower) ,Q3+1.5*IQR=80.5** (upper)

Which indicates there is no outliers.



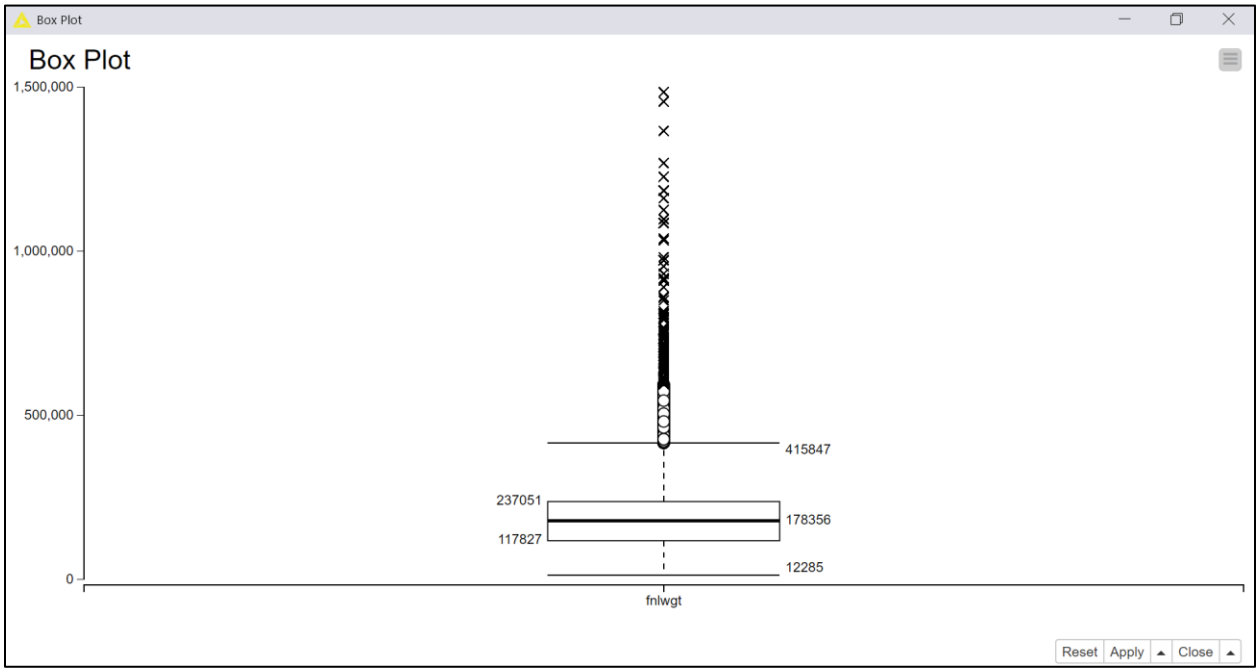
Age:

The boxplot shows outliers exists above value=78.



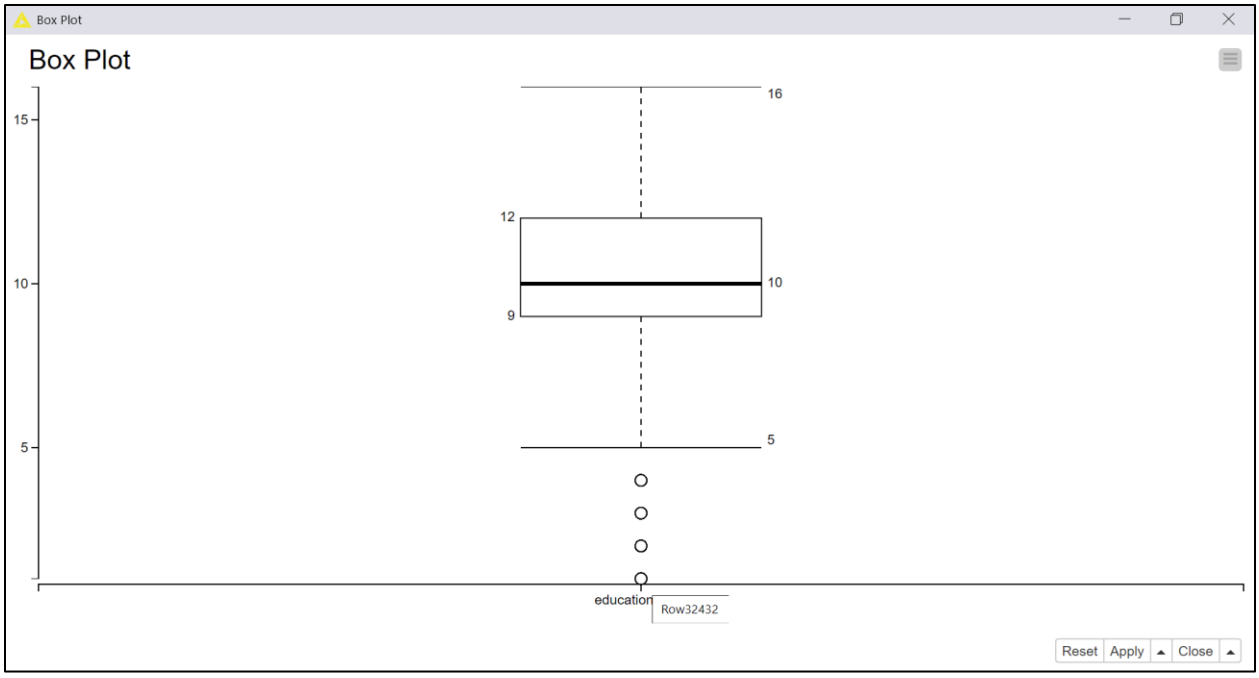
Fnlwgt

There are outlier exists for the values above= 594,683.



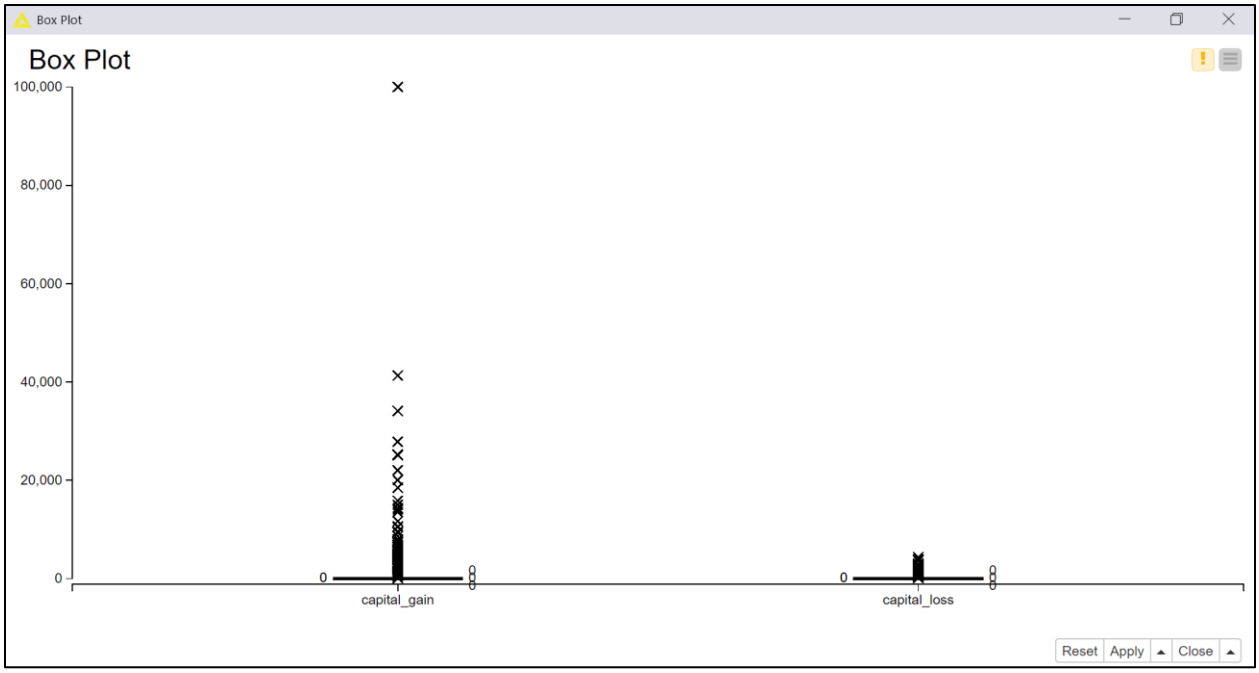
Education_num:

It has one outlier which is row number= 32432



Capital_gain and capital_loss

Capital_gain and capital_loss box plot doesn't show clearly although histogram showed outliers which were very visible.



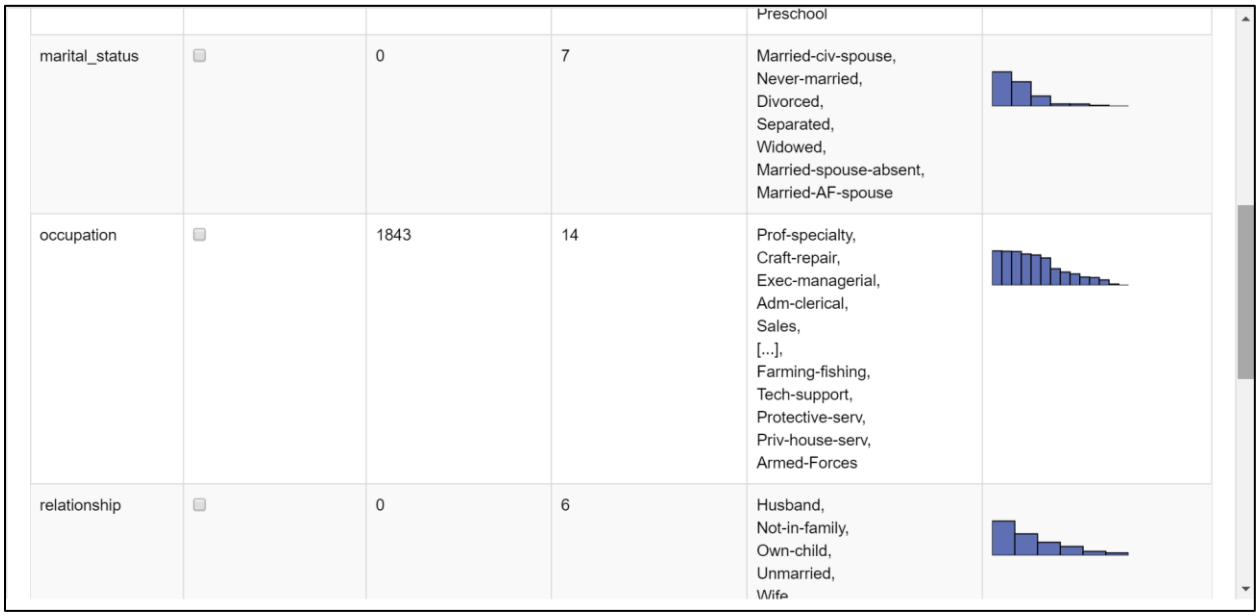
Categorical Variable:

The univariate analysis of the categorical variables is done by showing frequency distribution for each category in attribute using bar chart (Ray, 2016) (Chet Lemon, n.d.).

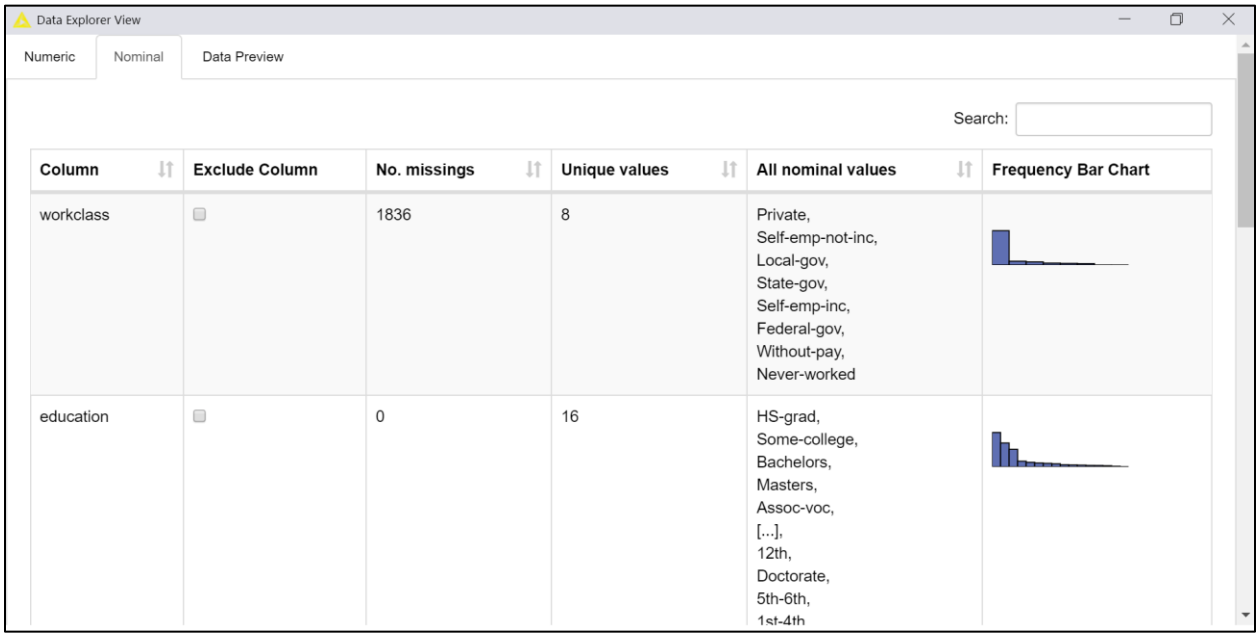
i) Data Explore Node:

Summary:

1. Workclass, occupation and native_country have missing values as can be seen from the figure below.
2. Native_country have highest number of distinct values 41 but data is not distributed equally. So this attribute doesn't give correct information.
3. Education and Marital_status uniques values can be reduced.



				Preschool	
marital_status	☐	0	7	Married-civ-spouse, Never-married, Divorced, Separated, Widowed, Married-spouse-absent, Married-AF-spouse	
occupation	☐	1843	14	Prof-specialty, Craft-repair, Exec-managerial, Adm-clerical, Sales, [...], Farming-fishing, Tech-support, Protective-serv, Priv-house-serv, Armed-Forces	
relationship	☐	0	6	Husband, Not-in-family, Own-child, Unmarried, Wife	



Data Explorer View

Numeric | Nominal | Data Preview

Search:

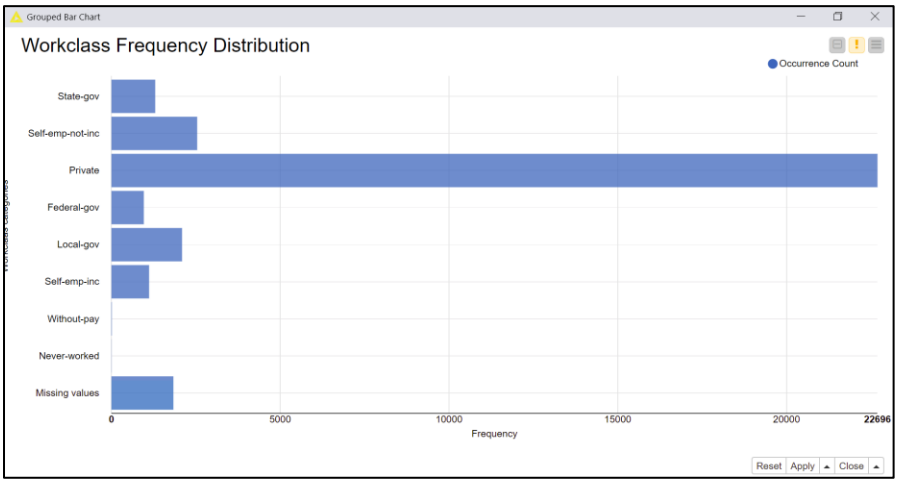
Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
workclass	☐	1836	8	Private, Self-emp-not-inc, Local-gov, State-gov, Self-emp-inc, Federal-gov, Without-pay, Never-worked	
education	☐	0	16	HS-grad, Some-college, Bachelors, Masters, Assoc-voc, [...], 12th, Doctorate, 5th-6th, 1st-4th	

				Other-relative	
race	☐	0	5	White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other	
sex	☐	0	2	Male, Female	
native_country	☐	583	41	United-States, Mexico, Philippines, Germany, Canada, [...], Outlying-US(Guam-USVI-etc), Hungary, Honduras, Scotland, Holand-Netherlands	
income	☐	0	2	<=50K, >50K	

ii) Frequency Distribution

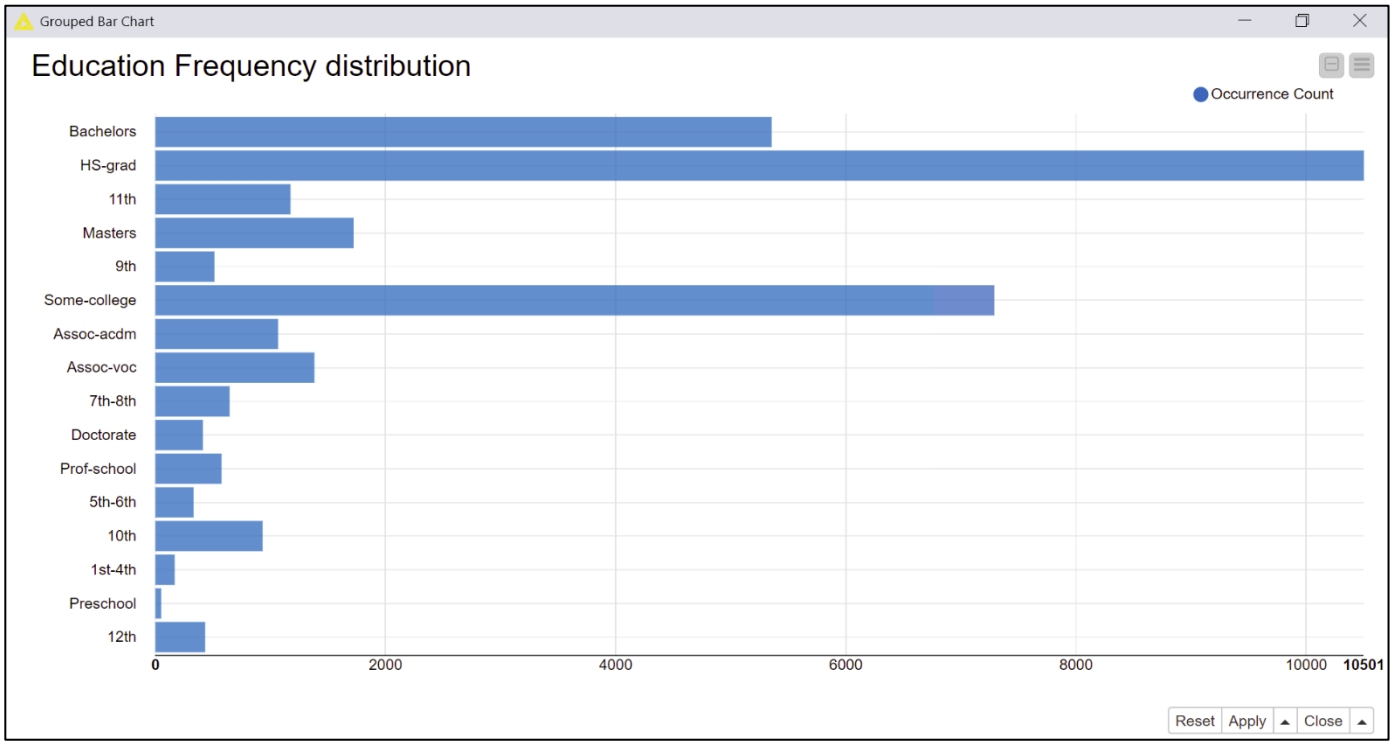
1. Workclass:

The **bar chart Node** shows the frequency distribution. It can be clearly seen 22696 records are doing jobs in private sector. There are almost zero people who dont work and few work without salary. The second most frequency distribution is self-employed-not-incorporated which means people are self employed but they don't company.



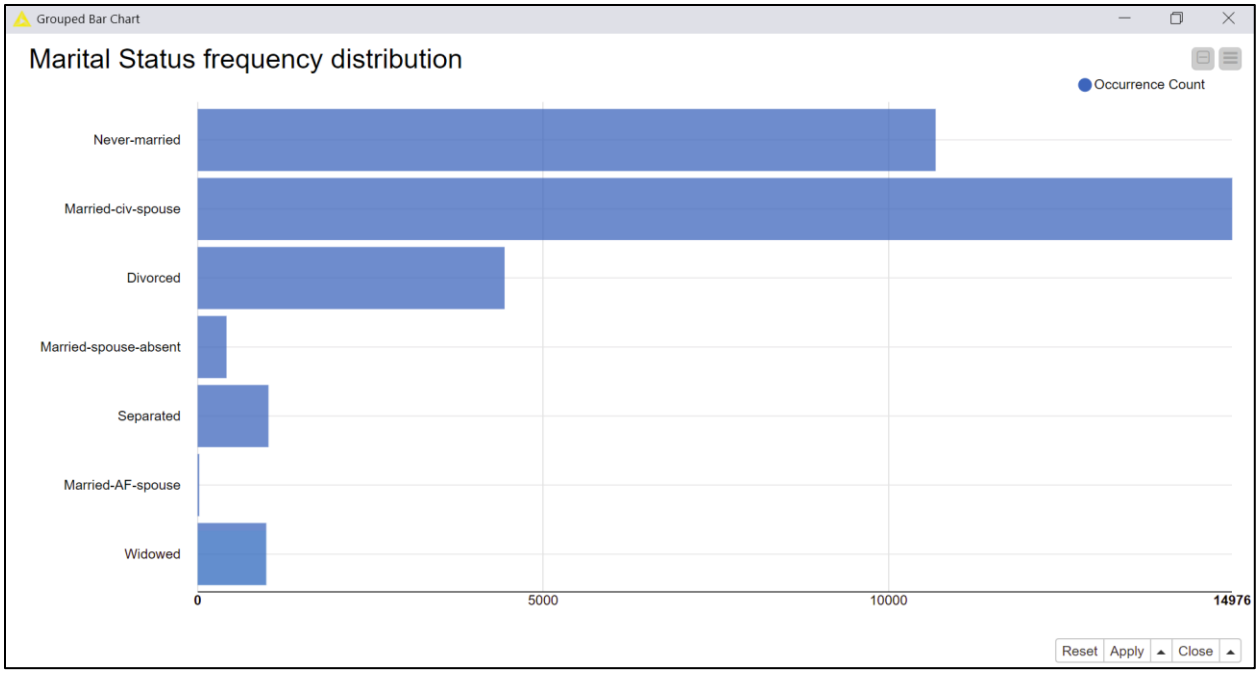
2. Education:

Barchart shows highest number of people which is 10501 people are High school in the dataset. The second highest education is some-college. Preschool has lowest frequency distribution. Education have 15 distinct values.



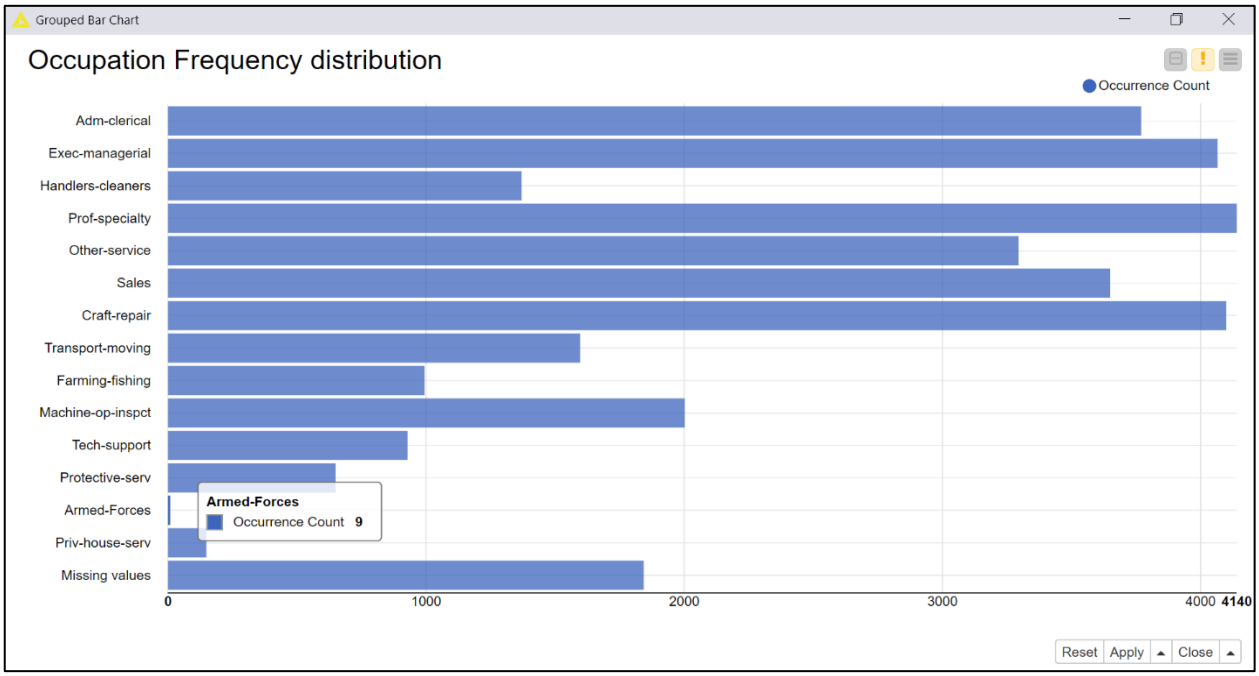
3. **Marital_status:**

- 14976 records in the dataset marital status is Married-civ-spouse which is married civilian spouse.
- Lowest people marital status is Married-AF-spouse which is Married Armed Forces Spouse.
- More than 10, 000 records which second highest records are never married.
- The marital_status have 7 different values.



4. Occupation:

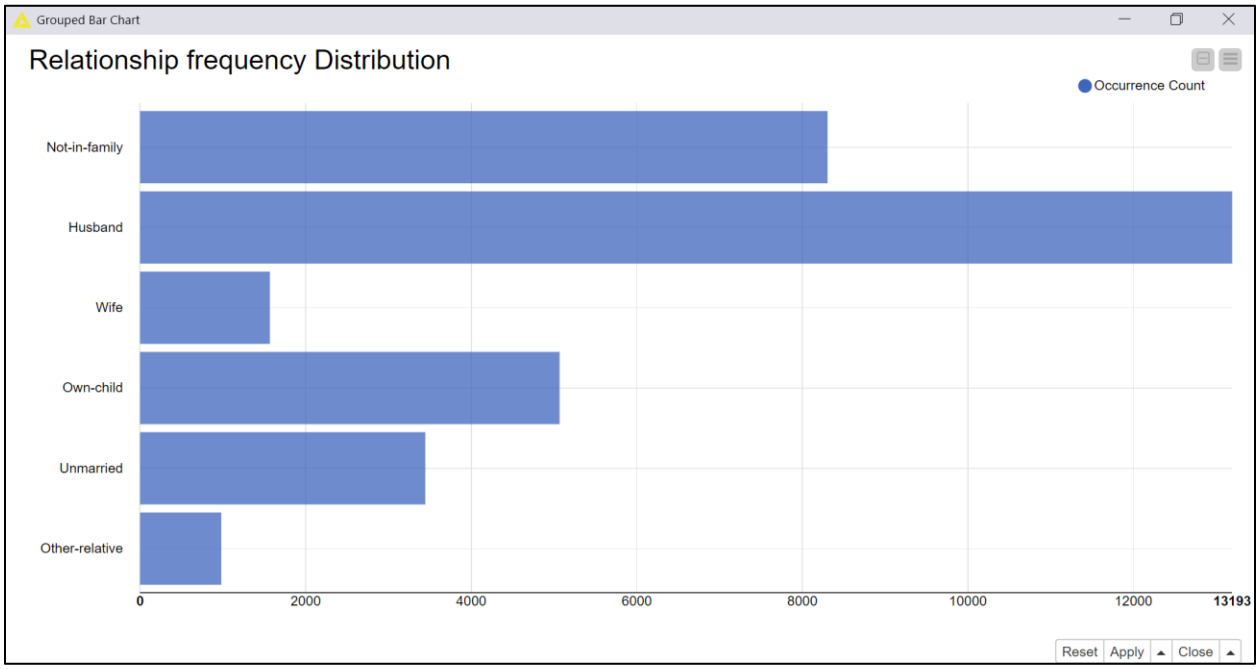
- The bar chart shows approximately normal distribution for all occupations except Armed-forces which is only present for 9 people. Highest number of people 4140 work as prof-specialty.
- Almost 2000 records donot have any occupation which means it has missing values.
- The occupation attribute have 14 different occupations.



5. Relationship:

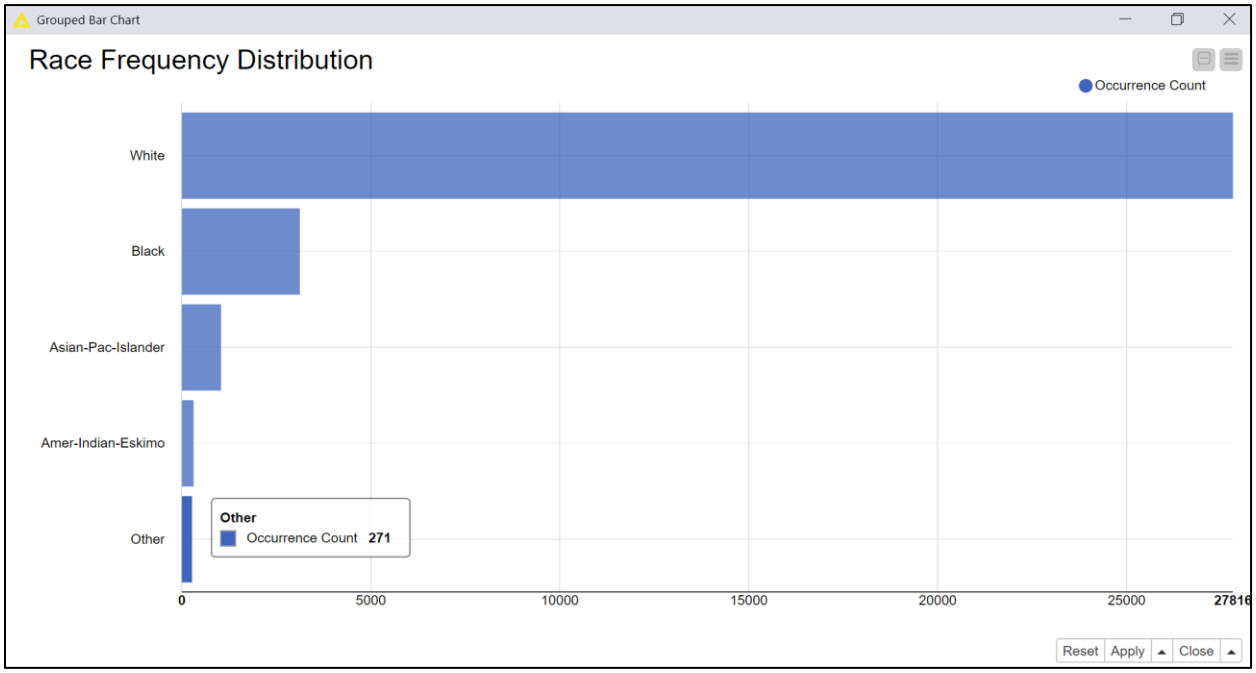
- Relationship doesn't have any missing values.

- The higest number of records relationship status is Husband which is 13193 records.
- Second highest records are for Not-in-family which is more than 8000 whereas other-relative is lowest status within the dataset.



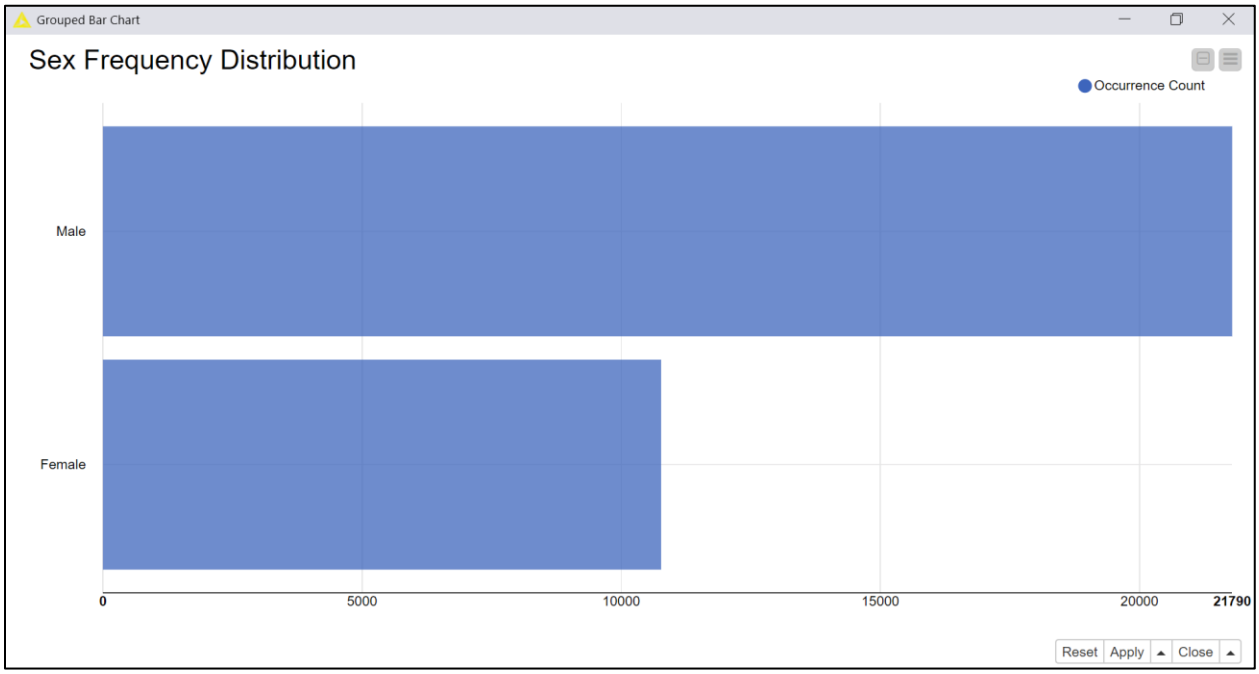
6. Race

- Most of the records ethnic origin is white which has total of records=27816.
- Lowest ethnicity is other with 271 records only.



7. **Sex**

The Sex which is gender has unequal distribution. 21790 records are male whereas rest are female.



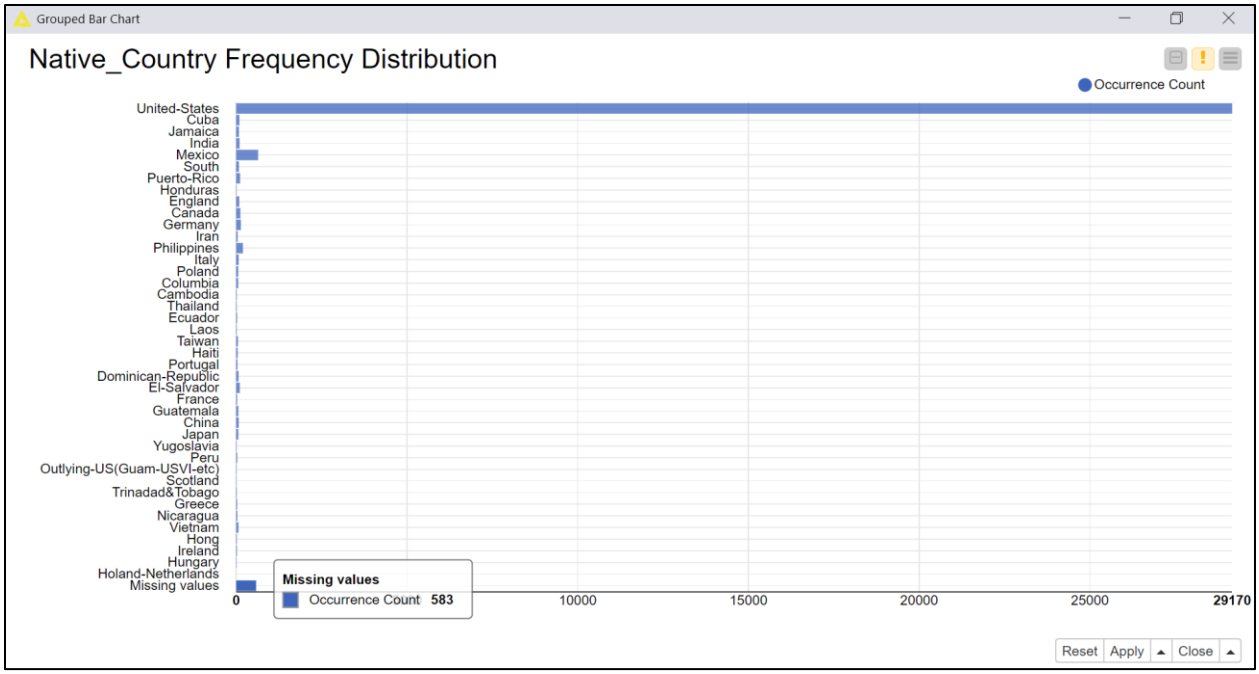
8. **Native_country:**

1.The barchart clearly show there is no normal distribution for the native_country of people.

2. 29170 people belong to United-states

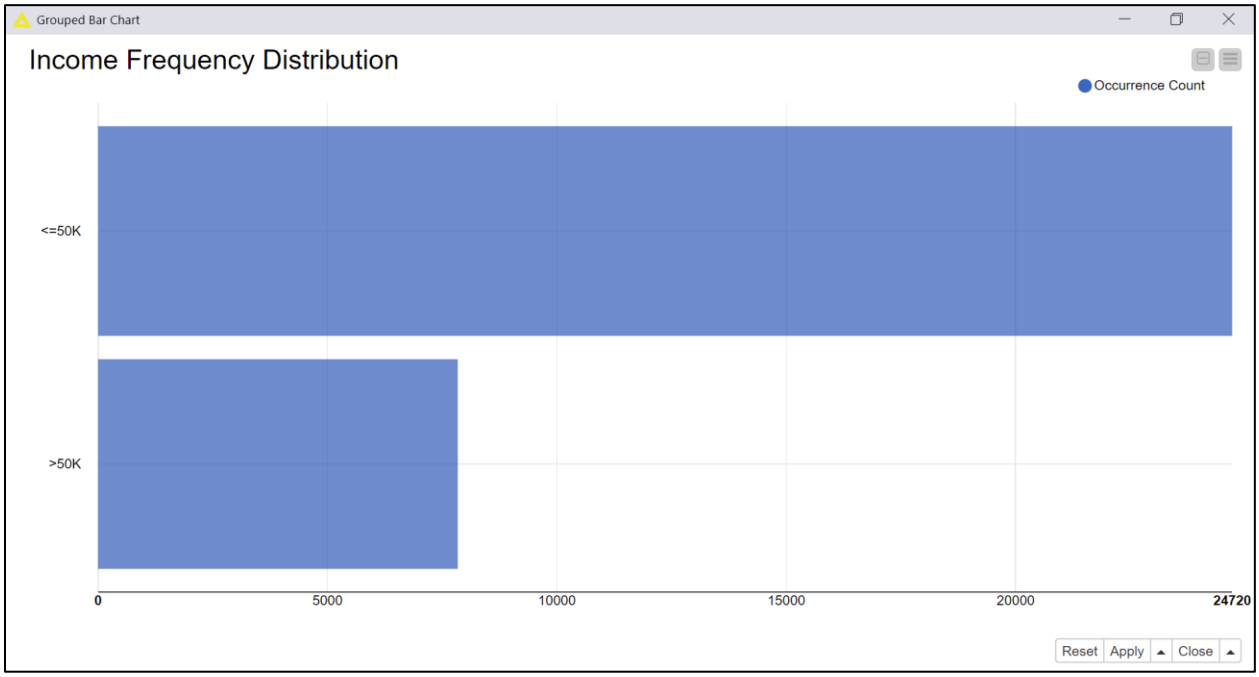
3. There are very small number of records from other countries so this attribute cant be used in analysis as it would give equal distribution to all values.

4. There are 583 total missing values.



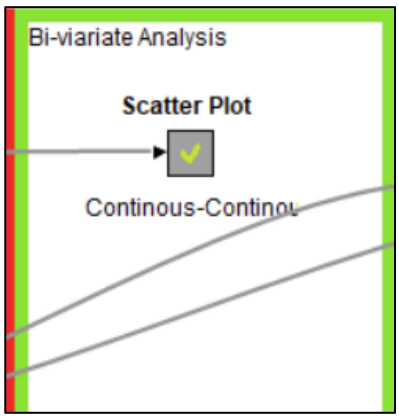
9. Income

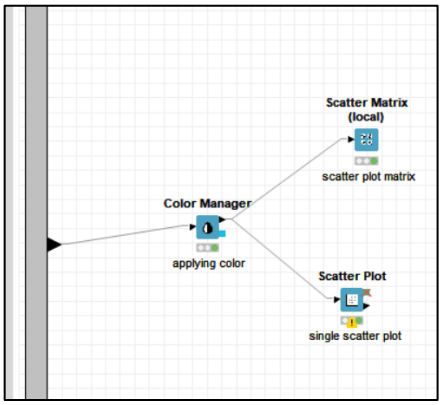
- 24720 people in record have salary <=50k
- Approximately 15000 records have salary >=50k
- Although this variable is not distributed normally but it is important to take into account.



4. Bi-Variate analysis

i) Continuous- continuous:





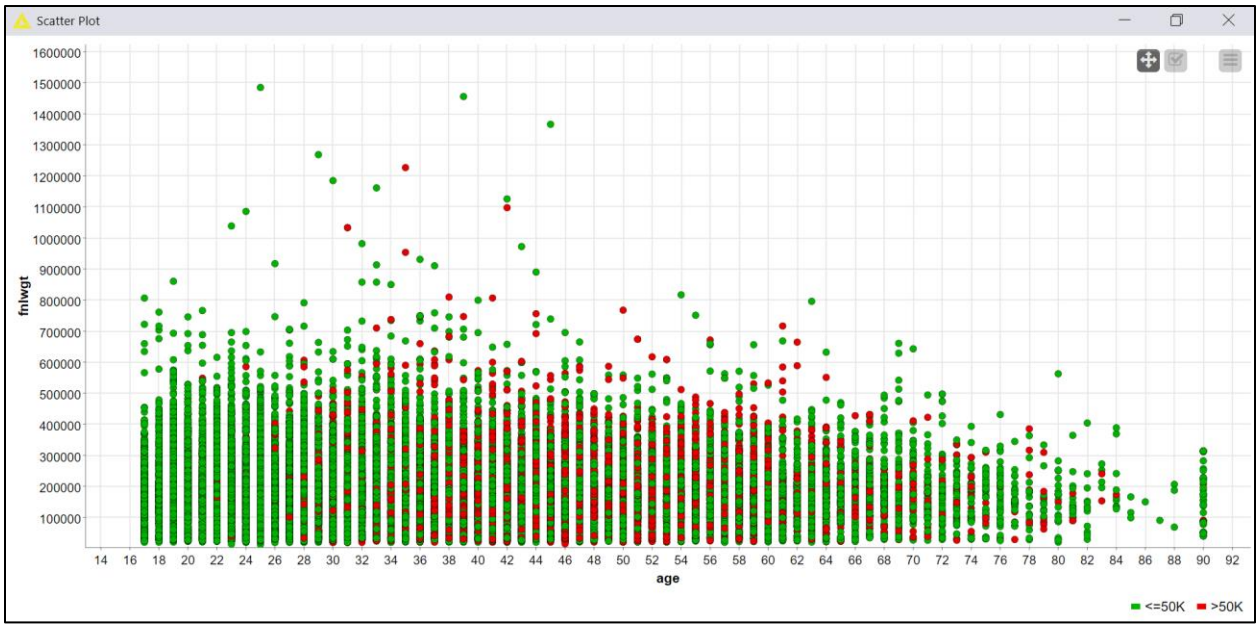
Scatterplot Meta Node is used to show relationship between two continuous variables. **Expanded MetaNode** has been showed on the right side which has **color manager** to apply color for categorical variable. X-axis and y-axis for numeric variable. **Scatter Matrix** shows all scatterplot for all variable in a matrix. **Scatter Plot Node** shows one scatter plot between two numeric variables only.

1. Age-fnlwgt

The scatterplot shows a little bit pattern of age with fnlwgt as it can be seen at younger age the fnlwgt has higher value as compared to after age 40 which can indicate the young people living in good residence and good employment. Age-fnlwgt shows opposite pattern.

Color is used to depict the salaries of people and fnlwgt attribute affect. However, the salaries at the young age are observed lower as compared to later age.

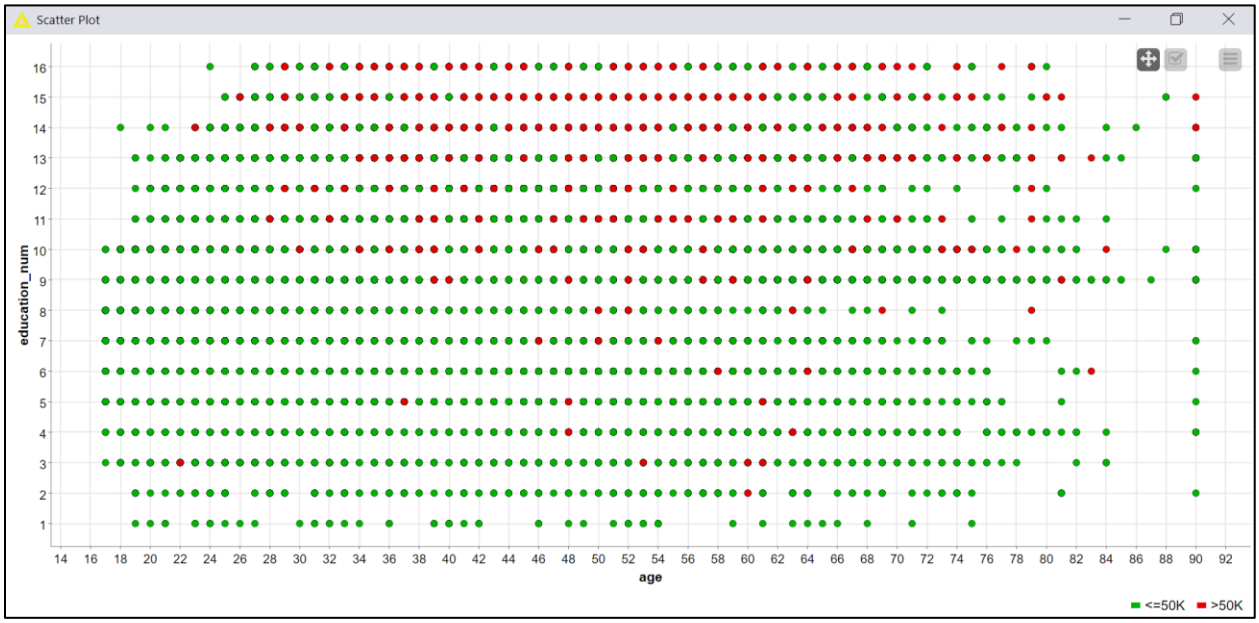
There are some outliers which are showing very high value for the fnlwgt attribute



2. Age-education_num

There is no pattern between age and education_num which is the educaton level. There is almost equal distribution for all education level at all ages.

However, color which is used to represent salary. The red higher salaries are observed in the population who are educated atleast 10 level. Below 10 education level, there are very few people have high salary than 50K.

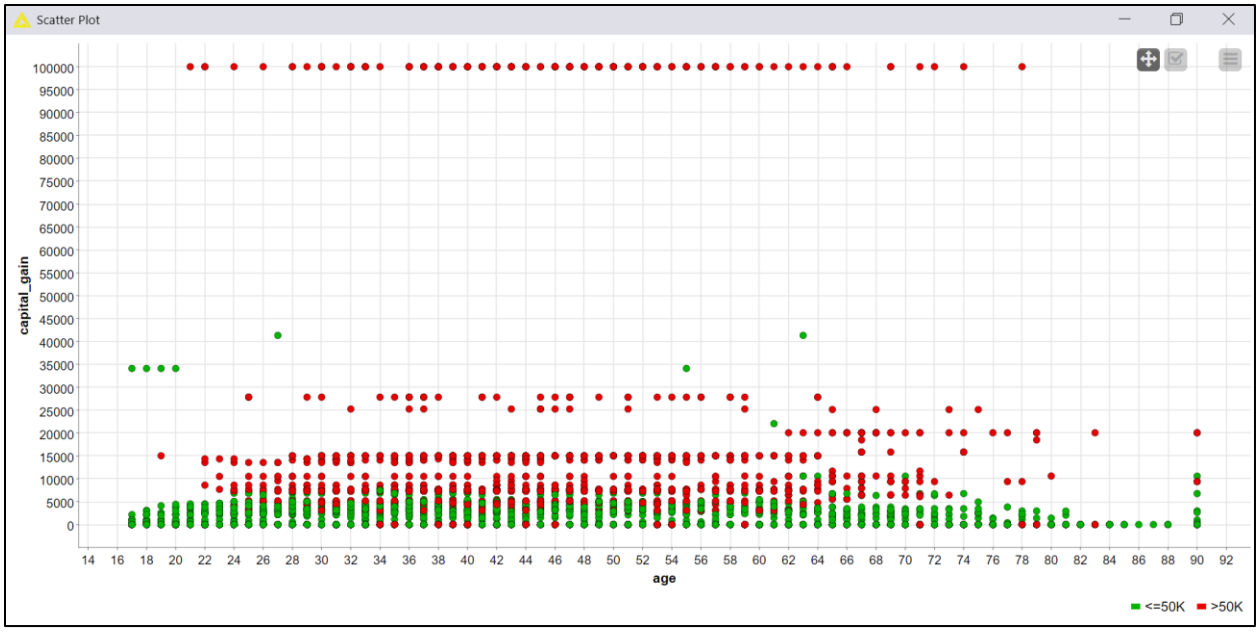


3. Age-capital_gain

1.There is no relationship between age and capital gain of an individual. It has equal distribution for all ages.

2. There is strange pattern of outliers can be seen in the capital_gain which has value 100000 after 35000. There is no value in the middle.

3. It can be seen by color, capital gain at all ages for salary <=50K has lower value whereas the people with higher salary have very high capital_gain.



Age-Capital_loss

Age and capital loss don't have any relationship.

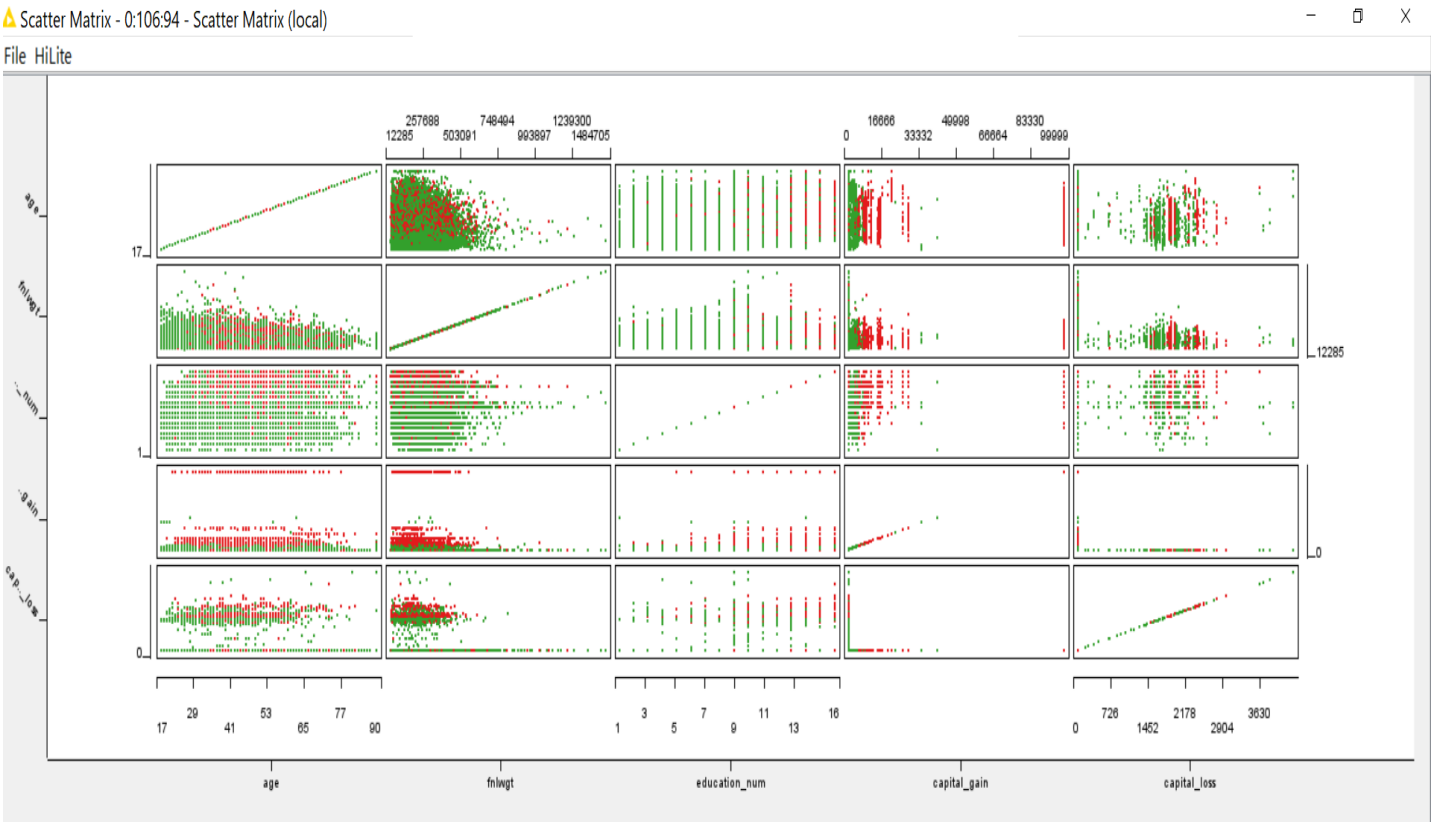
But, salary >-50k which is in red color shows that high capital loss because of high salary which is contradiction to capital_gain where it was increasing with salary increase.

Fnlwgt relationship:

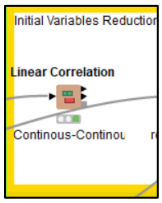
Capital_loss: Low value of fnlwgt, the higher the loss as. Salary>=50K can affect to increase the capital_loss as can be seen in the plot.

Capital_gain: For low values of the fnlwgt there is high capital_gain. The people with high salary>=50k shows more capital_gain.

Age: Increase in age has negative affect decreasing the fnlwgt as can be seen in scatter plot.



Correlation Matrix(continuous-categorical):



i) **Workclass** have strong correlation which is 0.4 with **occupation**, because they are giving almost same information which is repeating. One attribute can be removed to avoid redundant information.

ii) Cross in the boxes shows numeric correlation cannot exist with categorical.

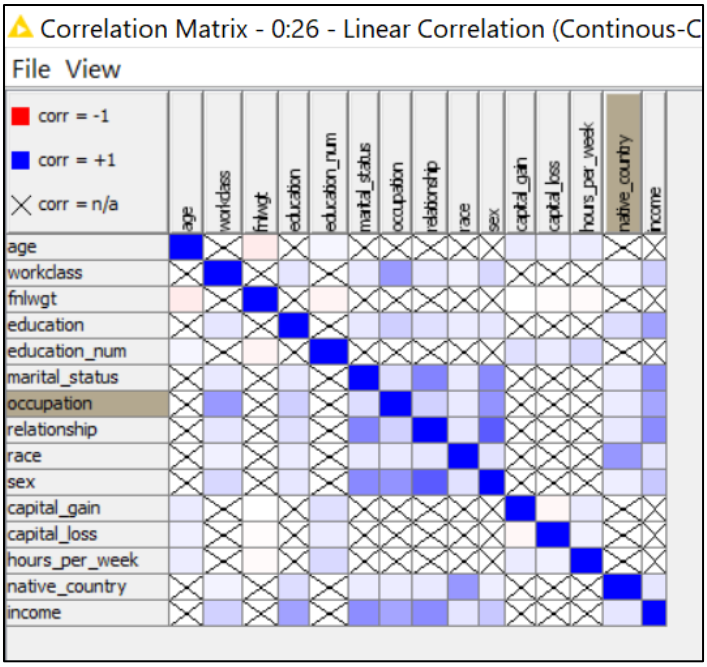
iii) **Education** have 0.36 correlation with **income**.

iv) **Marital status** have correlation with **relationship, sex, income** which is more than 0.4.

v) **Occupation have correlation with workclass, sex, income.**

vi) **Race** have correlation with **native_country** which is more than 0.4.

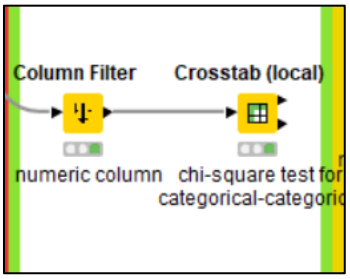
vii) Looking at the correlation**, workclass,sex,income and native_country** features can be removed to avoid duplicate information.



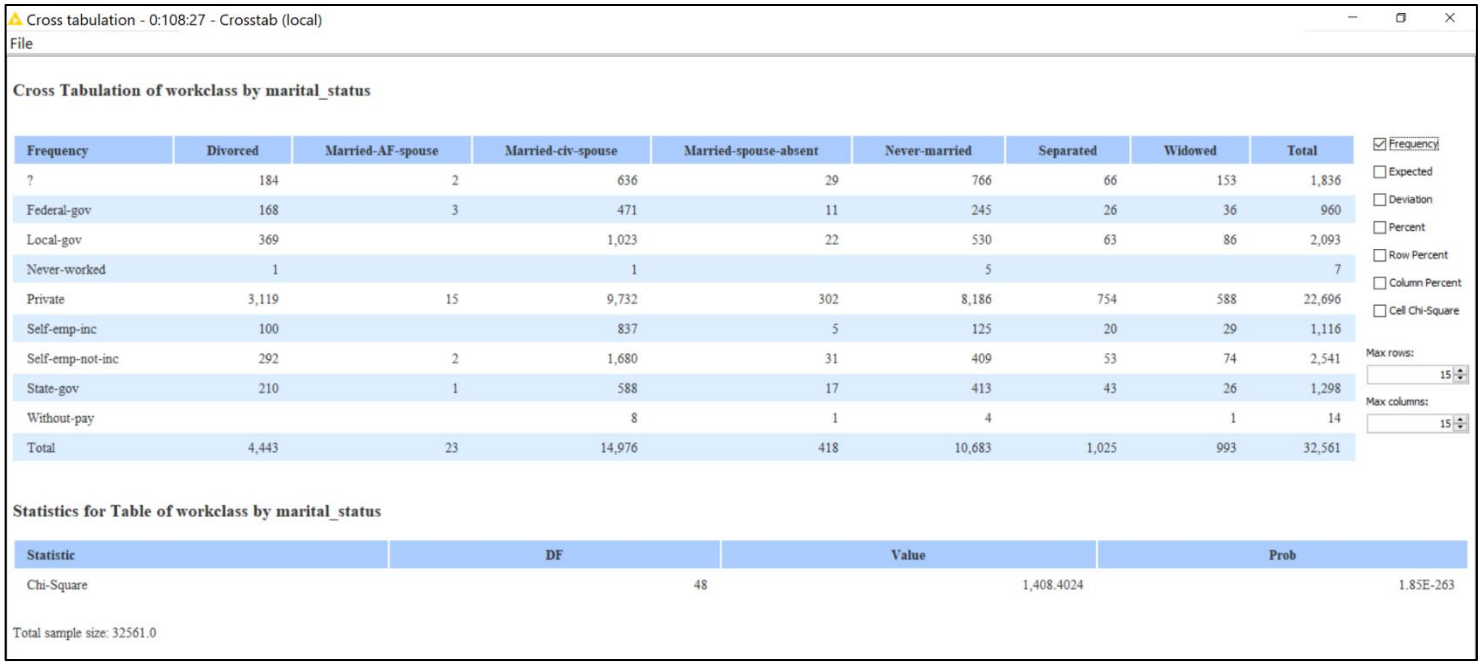
Categorical-categorical

Cross tabulation(Chi-square Test):

The cross tabulation shows the frequency of values for each category of workclass with each category of marital_status. The value of probabailty $p<0.05$ means there is significant association between the categorical attributes[2] (Larose, 2005)**.** Here p=1.85e-263 which means $p < 0.05$ which shows 95% of significance of association between the **workclass and marital_status.**



[2] https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/#one



Cross tabulation - 0:108:27 - Crosstab (local)

File

Cross Tabulation of workclass by marital_status

Frequency	Divorced	Married-AF-spouse	Married-civ-spouse	Married-spouse-absent	Never-married	Separated	Widowed	Total
?	184	2	636	29	766	66	153	1,836
Federal-gov	168	3	471	11	245	26	36	960
Local-gov	369		1,023	22	530	63	86	2,093
Never-worked	1		1		5			7
Private	3,119	15	9,732	302	8,186	754	588	22,696
Self-emp-inc	100		837	5	125	20	29	1,116
Self-emp-not-inc	292	2	1,680	31	409	53	74	2,541
State-gov	210	1	588	17	413	43	26	1,298
Without-pay			8	1	4		1	14
Total	4,443	23	14,976	418	10,683	1,025	993	32,561

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square
Max rows: 15
Max columns: 15

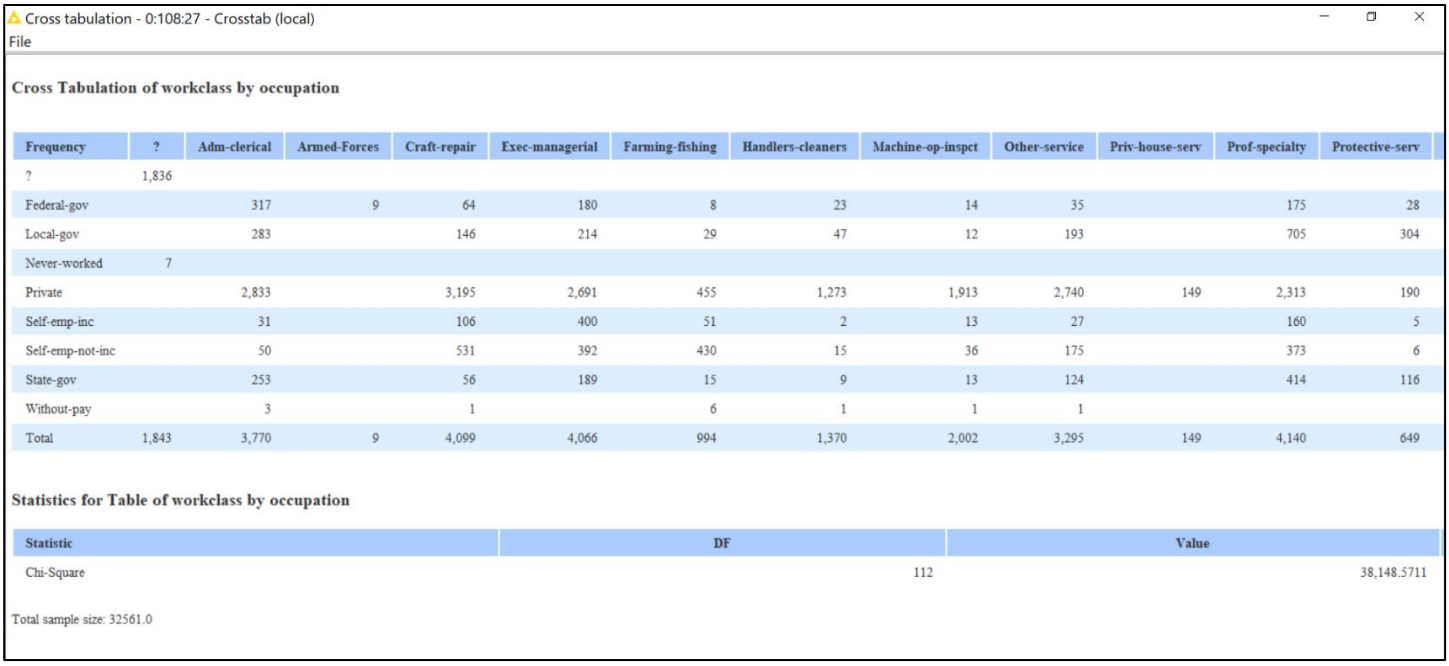
Statistics for Table of workclass by marital_status

Statistic	DF	Value	Prob
Chi-Square	48	1,408.4024	1.85E-263

Total sample size: 32561.0

Workclass- Occupation:

- 1836 are missing values with ? mark.
- Probability is zero which means there is 95% association between the attributes.



Cross tabulation - 0:108:27 - Crosstab (local)

File

Cross Tabulation of workclass by occupation

Frequency	?	Adm-clerical	Armed-Forces	Craft-repair	Exec-managerial	Farming-fishing	Handlers-cleaners	Machine-op-inspct	Other-service	Priv-house-serv	Prof-specialty	Protective-serv
?	1,836											
Federal-gov		317	9	64	180	8	23	14	35		175	28
Local-gov		283		146	214	29	47	12	193		705	304
Never-worked	7											
Private		2,833		3,195	2,691	455	1,273	1,913	2,740	149	2,313	190
Self-emp-inc		31		106	400	51	2	13	27		160	5
Self-emp-not-inc		50		531	392	430	15	36	175		373	6
State-gov		253		56	189	15	9	13	124		414	116
Without-pay		3		1		6	1	1	1			
Total	1,843	3,770	9	4,099	4,066	994	1,370	2,002	3,295	149	4,140	649

Statistics for Table of workclass by occupation

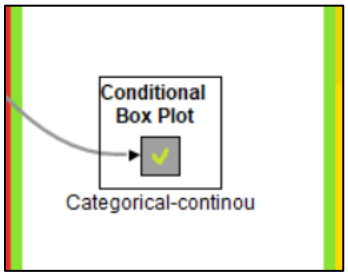
Statistic	DF	Value
Chi-Square	112	38,148.5711

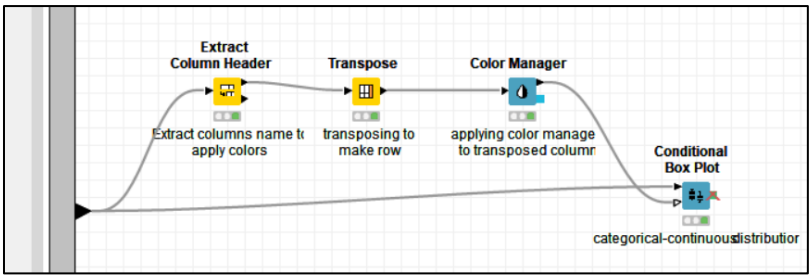
Total sample size: 32561.0

Categorical-continuous:

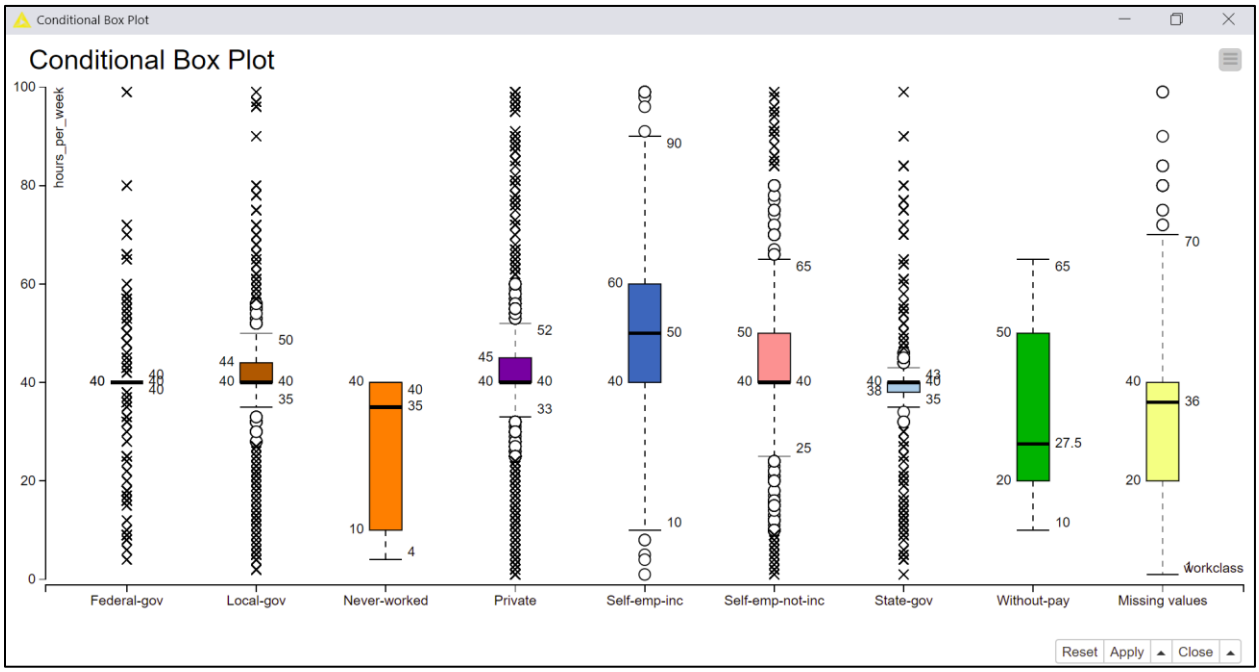
Conditional Boxplot:

The Conditional boxplot node has been used. The conditional box plot allows to look at relationship between categorical and continuous variables. The conditional boxplot node has been expanded in the following diagram.





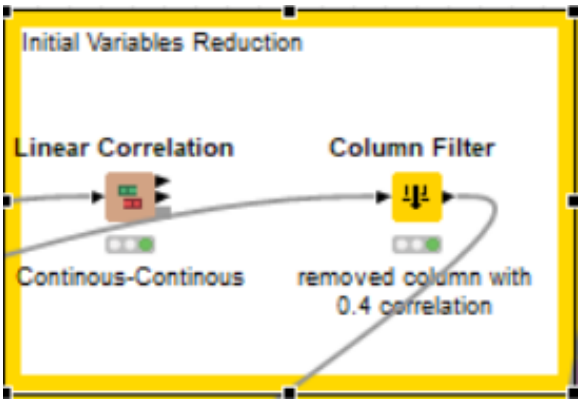
The Extract Column Header node will extract column names and store them in a row. **Transpose node** will convert them to one column. **Color Manager** node will apply different color to the value of this column. The conditional box plot below shows workclass all categories on x-axis which is categorical attribute. The y-axis shows **hours_per_week** which is continuous variable. Color is representing different categories of **Workclass**. This conditional box plot shows how hours are distributed in each category as **Self-emp-incorporated** have highest hours =90.



Preparation of Data:

1. Removing Variable with Redundant Information

During initial analysis and visualization of variables, insignificant or attributes which contain same/duplication information can be removed to avoid unnecessary transformation on variables. **Linear Correlation Node** has been used to identify strong relationship between variables as explained previously and then used **Correlation Filter Node** to set correlation threshold=0.4 to filter out variables with 0.4 positive correlation.



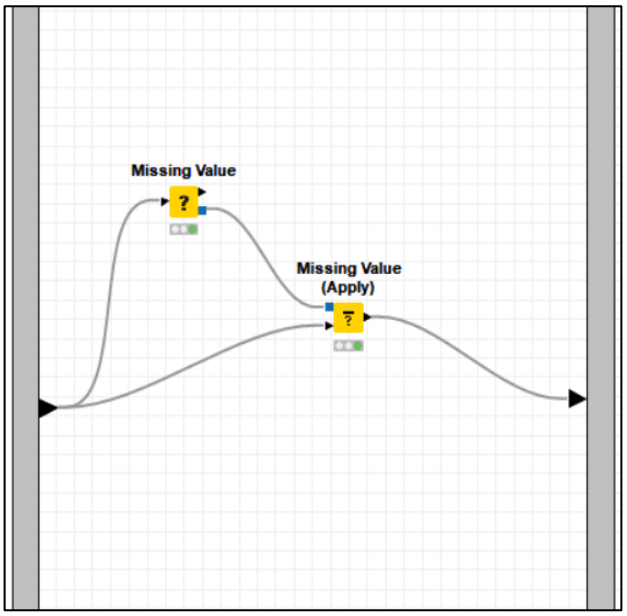
The relationship, sex, native_country and marital_status columns have been removed as they had strong correlation with other variables which was 0.4.

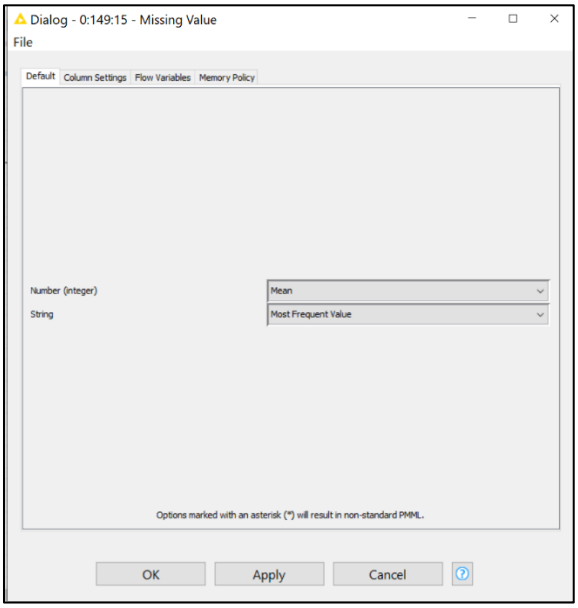
2. Missing Values treatment

There are three categorical attributes which have missing values present in the data. The following columns have missing values.

workclass	1836
occupation	1843
native_country	583

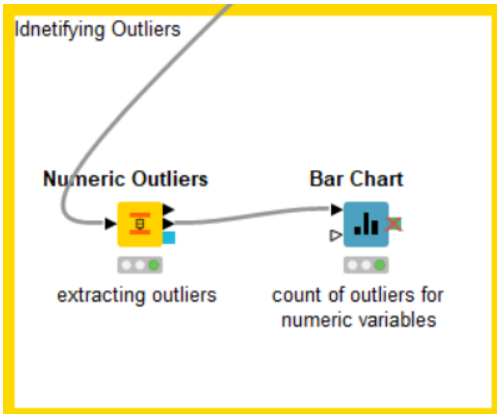
But, **occupation** and **native_country** has been removed from the analysis because of strong correlation with other variables in first step. Therefore, there is need to treat only one column values which is **workclass**. Replacing the missing values with frequent value as shown in the screen below.





3. Outliers Detection and Treatment:

During univariate and bivariate analysis outliers were identified using Boxplot, Histogram and scatterplot. Here, the outliers are not removed because they are part of natural pattern in the data. Here, **Numeric binner Node is only used to list** the number of outliers in the table form for numeric variables only.



The figure shows that all six numeric attributes have outliers. Hours_per_week have highest number of outliers as 9008.

Summary - 0:158 - Numeric Outliers

File Hilite Navigation View

Table "default" - Rows: 6 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Outlier ...	Member count	Outlier count	Lower bound	Upper bound
Row0	age	32561	143	-2	78
Row1	fnlwgt	32561	992	-61,027	415,895
Row2	education_n...	32561	1198	4.5	16.5
Row3	capital_gain	32561	2712	0	0
Row4	capital_loss	32561	1519	0	0
Row5	hours_per_...	32561	9008	32.5	52.5

Figure 1: Outlier count

No observation will be removed to deal with outliers as the outlier looks like the part of pattern. Log transformation and Normalization will be applied in subsequent steps to deal with outliers and to make them normally distributed.

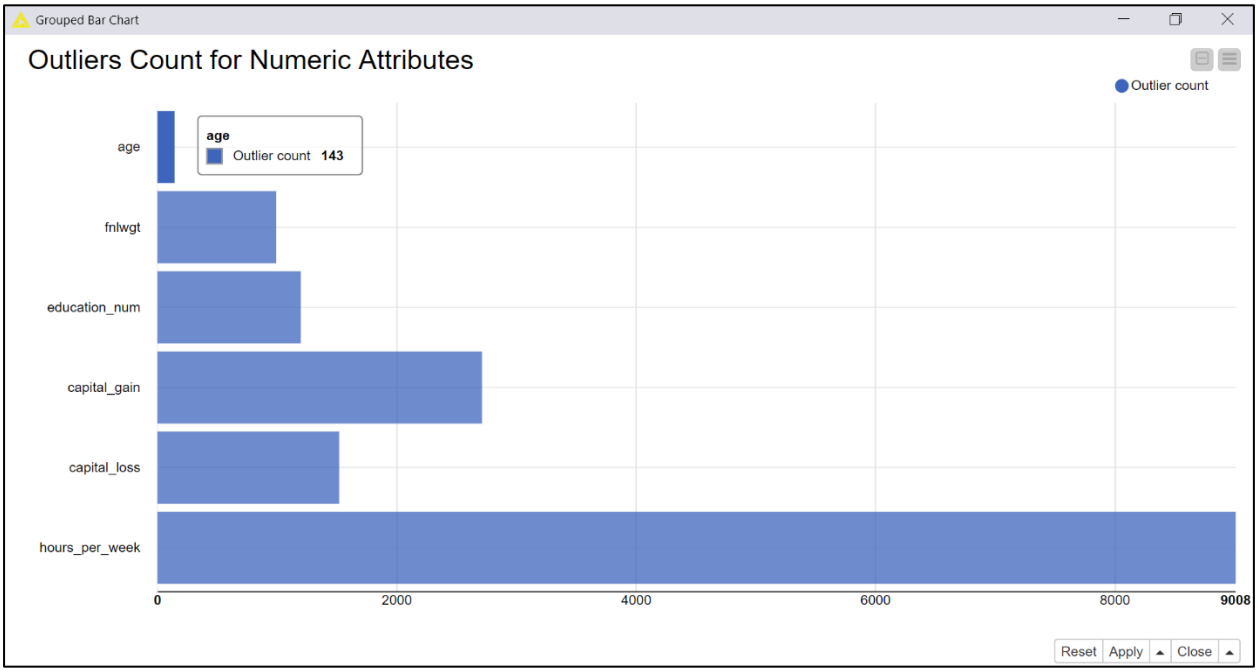
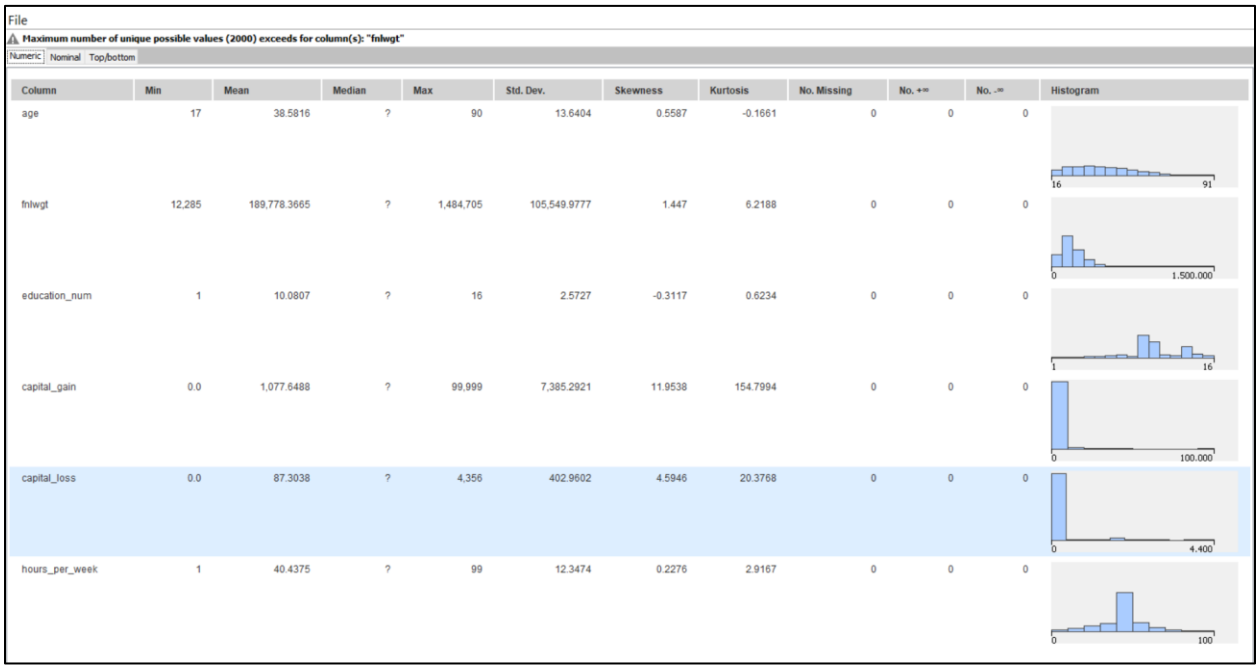


Figure 2: outlier bar chart

4. Normalization

Normalization is the process of transforming the variable into new range using some function. Normalization is important for two reasons. First, if the range of the values for the attributes is different, then it can influence in the analysis giving more importance to one variable than other. Second, some data mining algorithm require to apply normalization as these models cannot run without normalizing the data (Myatt, 2007).

In outlier identification section, it was observed numeric variables contain the outliers. These variables are not normally distributed except **hours_per_week** which is not good for any data mining algorithm. **Fnlwgt, capital_loss, capital gain and age** are positively skewed and square root, cube and log transformation can be applied. **Education_num** is negatively skewed/right skewed. The following screen shows the histogram and distribution.

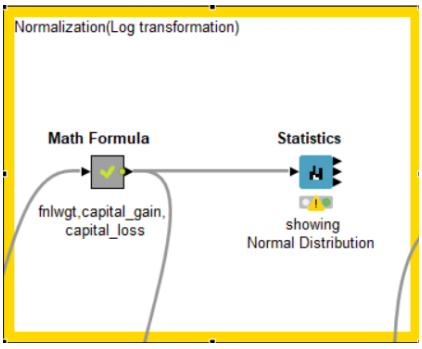

File

Maximum number of unique possible values (2000) exceeds for column(s): "fnlwgt"

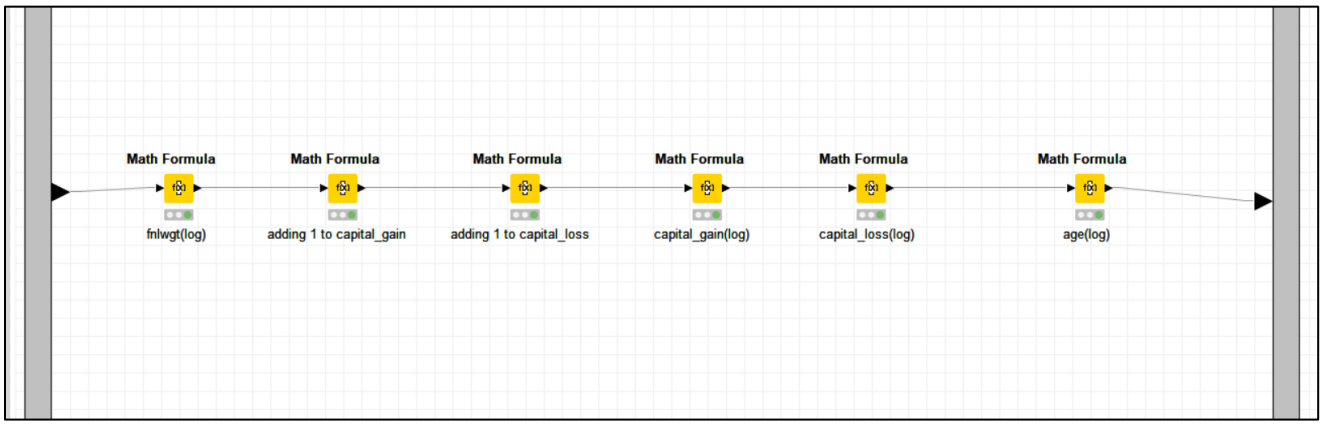
Numeric | Nominal | Top/bottom

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
age	17	38.5816	?	90	13.6404	0.5587	-0.1661	0	0	0	16 … 91
fnlwgt	12,285	189,778.3665	?	1,484,705	105,549.9777	1.447	6.2188	0	0	0	0 … 1,500,000
education_num	1	10.0807	?	16	2.5727	-0.3117	0.6234	0	0	0	1 … 16
capital_gain	0.0	1,077.6488	?	99,999	7,385.2921	11.9538	154.7994	0	0	0	0 … 100,000
capital_loss	0.0	87.3038	?	4,356	402.9602	4.5946	20.3768	0	0	0	0 … 4,400
hours_per_week	1	40.4375	?	99	12.3474	0.2276	2.9167	0	0	0	0 … 100

The following knime flow shows one **Metanode for Math formula** which is used to apply log transformation.



The following screen shows expanded Metanode subnodes which are transforming four variables.

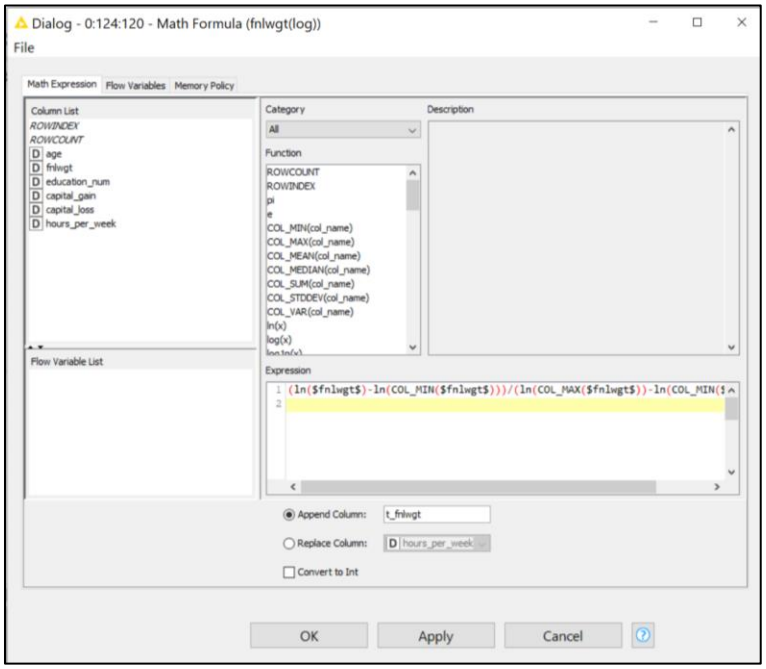


Capital_gain and Capital_loss had zero as value and it is not possible to apply log transformation if a column contain zero values. To overcome this problem, 1 is added to capital_gain and capital_loss columns and then applied the log transformation formula. Log, square root and cube transformation are applied to convert column into normal distribution. Knime doesn't offer any dedicated node for all these transformations. So, Math Formula will be used to apply log transformation which provides high accuracy for transforming the data.

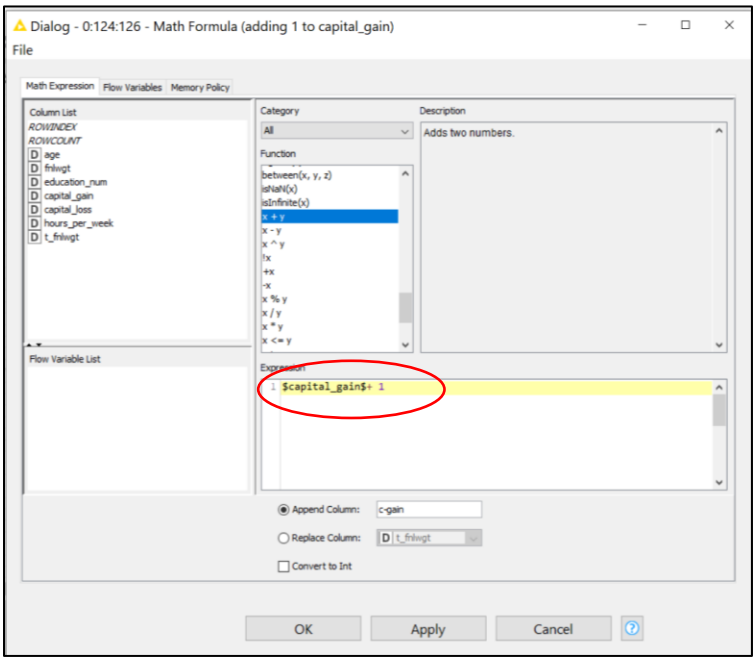
The following formula has been used in the Math Formula Node.

$$f_{\ln}(v) = \frac{\ln(v) - \ln(\min)}{\ln(\max) - \ln(\min)}$$

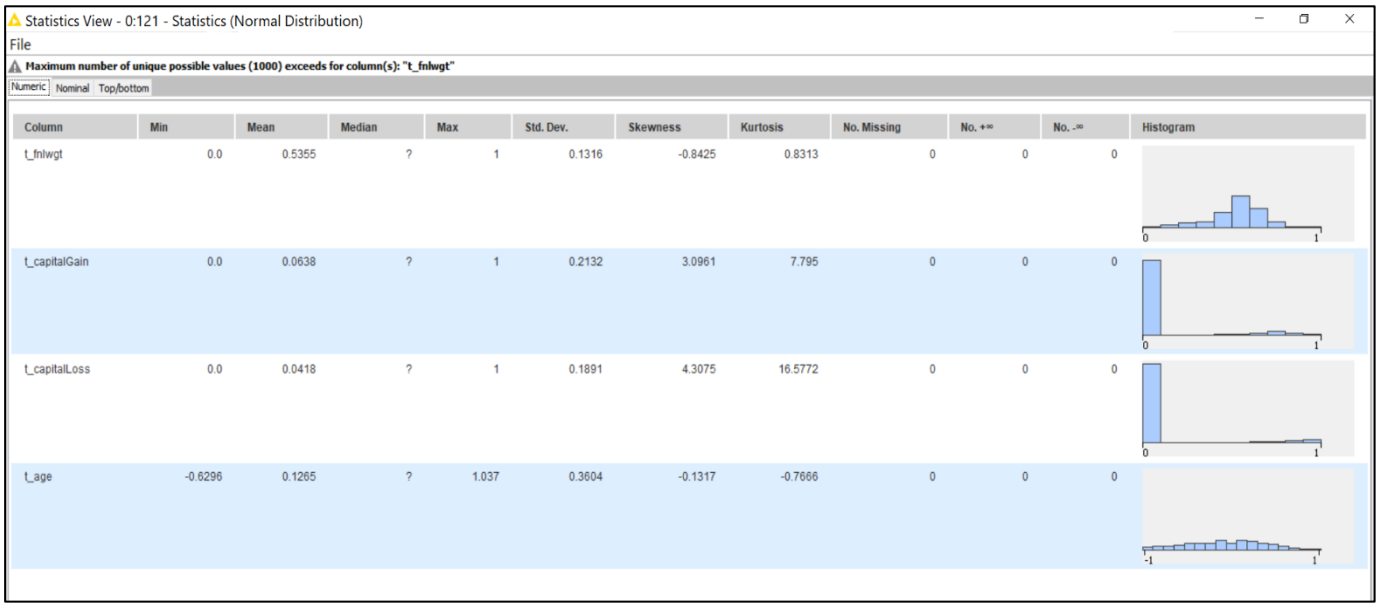
The following screen shows formula applied in the **Math Formula Node.**



The variables **capital_gain** and **capital_loss** have value 0 in the column. To apply log transformation to these attributes, there is needed to add 1 to all values of the column as it can be seen in the screen below.



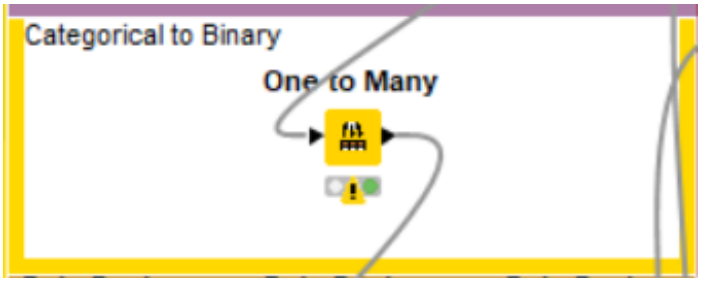
After applying log transformation to **age,capital_loss, capital_gain and fnlwgt**, the following screen shows distribution. The distribution is not perfect, but it got better. Age,fnlwgt are now normally distributed but the affect on capital _loss and capital _gain is very small.



4. Categorical to Binary Conversion

To apply the clustering algorithm, the distance is calculated between all the attributes to find similarity and group the observations into the clusters. The dataset should be converted into

numeric to apply specific distance measure. In this step, **One to Many Node is used to** create multiple columns for all distinct values of a categorical attribute and assigned 0 or 1 value for columns.



The table below shows the categorical variables converted to binary columns. New columns are renamed ***value_original_column_name.***

Processed data - 0:69 - One to Many

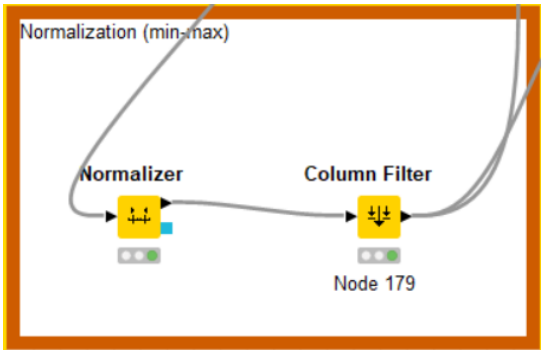
File Hilite Navigation View

Table "default" - Rows: 32561 Spec - Columns: 68 Properties Flow Variables

Row ID	State-g...	Self-em...	Private...	Federal...	Local-g...	Self-em...	Withou...	Never-...	missing...	Bachelo...	HS-gra...	11th_e...	Master...	9th_ed...	Some-c...
Row0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Row1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
Row2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row3	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row4	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Row5	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Row6	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
Row7	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
Row8	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Row9	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Row10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Row11	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Row12	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Row13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Row14	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Row15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Row16	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
Row17	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row18	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row19	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
Row20	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Row21	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row22	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
Row23	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row24	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row25	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
Row26	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row27	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Row28	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row29	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Row30	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Row31	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
Row32	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
Row33	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Row34	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Row35	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row36	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1

5. Min-Max Normalization:

Normalizer Node has been used to scale education_num and hour_per_week between 0 and 1. Log transformation applied to other variables converted them to 0-1 except **age** variable which is transformed to -1 to 1. So , applied min-max normalization as it shrinks extreme values to mean (Myatt, 2007).

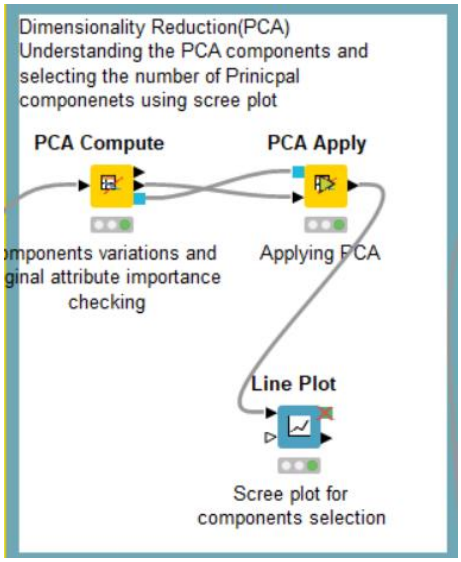


Dimensionality Reduction:

1. PCA understanding and components selection

PCA is technique which gives principal components attributes <= original attributes. Each principal component groups some attributes based on importance and gives weights to each attribute. The purpose of using PCA is select the principal components which have highest % of variations. The purpose of the principal component is that two components are uncorrelated. The first principal component always has highest variations. Looking at the eigenvector for the % variation each principal component have, the number of principal components has been selected (Glenn J. Myatt, 2009). From visual analysis, four attributes were removed which were presenting redundant information and were not significant to the domain knowledge.

The following knime workflow to understand the PCA and selecting the number of PCA components.



PCA Compute Node is used as it gives three outputs. The first table it gives is covariance matrix which displays correlation between all the original dimensions as shown below.

Covariance matrix - 0:139 - PCA Compute (Components variations and)

File Hilite Navigation View

Table "covariance matrix" - Rows: 24 Spec - Columns: 24 Properties Flow Variables

Row ID	educati...	hours_...	t_fnlwgt	t_capit...	t_capit...	t_age	State-gov	Self-em...	Private	Federal...	Local-gov	Self-em...	Withou...	Never-...	missing	White	Black	Asian-P...	Amer-I...
education_num	0.029	0.003	-0.001	0.005	0.003	0.003	0.003	0.001	-0.009	0.002	0.004	0.002	-0	-0	-0.003	0.003	-0.004	0.002	-0
hours_per_w...	0.003	0.016	-0	0.002	0.001	0.004	-0.001	0.003	-0.001	0	0	0.003	-0	-0	-0.005	0.002	-0.002	-0	-0
t_fnlwgt	-0.001	-0	0.017	-0	-0	-0.002	-0.001	-0.002	0.004	-0.001	0	-0	-0	0	-0	-0.002	0.004	-0.001	-0.001
t_capitalGain	0.005	0.002	-0	0.045	-0.003	0.006	-0	0.001	-0.004	0	0	0.003	0	-0	-0.001	0.002	-0.002	-0	-0
t_capitalLoss	0.003	0.001	-0	-0.003	0.036	0.002	-0	0.001	-0.002	0	0.001	0.001	-0	-0	-0.001	0.001	-0.001	0	-0
t_age	0.003	0.004	-0.002	0.006	0.002	0.047	0.001	0.008	-0.019	0.002	0.004	0.004	0	-0	-0	0.002	-0.001	-0	-0
State-gov	0.003	-0.001	-0.001	-0	-0	0.001	0.038	-0.003	-0.028	-0.001	-0.003	-0.001	-0	-0	-0.002	-0.001	0.001	0.001	0
Self-emp-not-inc	0.001	0.003	-0.002	0.001	0.001	0.008	-0.003	0.072	-0.054	-0.002	-0.005	-0.003	-0	-0	-0.004	0.005	-0.005	-0	-0
Private	-0.009	-0.001	0.004	-0.004	-0.002	-0.019	-0.028	-0.054	0.211	-0.021	-0.045	-0.024	-0	-0	-0.039	0	-0	-0	-0.001
Federal-gov	0.002	0	-0.001	0	0	0.002	-0.001	-0.002	-0.021	0.029	-0.002	-0.001	-0	-0	-0.002	-0.003	0.002	0	0
Local-gov	0.004	0	0	0	0.001	0.004	-0.003	-0.005	-0.045	-0.002	0.06	-0.002	-0	-0	-0.004	-0.002	0.003	-0.001	0
Self-emp-inc	0.002	0.003	-0	0.003	0.001	0.004	-0.001	-0.003	-0.024	-0.001	-0.002	0.033	-0	-0	-0.002	0.003	-0.003	0	-0
Without-pay	-0	-0	-0	0	-0	0	-0	-0	-0	-0	-0	-0	0	-0	-0	0	-0	0	-0
Never-worked	-0	-0	0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	0	-0	-0	0	-0	-0
missing	-0.003	-0.005	-0	-0.001	-0.001	-0	-0.002	-0.004	-0.039	-0.002	-0.004	-0.002	-0	-0	0.053	-0.002	0.001	0	0
White	0.003	0.002	-0.002	0.002	0.001	0.002	-0.001	0.005	0	-0.003	-0.002	0.003	0	-0	-0.002	0.124	-0.082	-0.027	-0.008
Black	-0.004	-0.002	0.004	-0.002	-0.001	-0.001	0.001	-0.005	-0	0.002	0.003	-0.003	-0	0	0.001	-0.082	0.087	-0.003	-0.001
Asian-Pac-Isl...	0.002	-0	-0.001	-0	0	-0	0.001	-0	-0	0	-0.001	0	0	-0	0	-0.027	-0.003	0.031	-0
Amer-Indian-...	-0	-0	-0.001	-0	-0	-0	0	-0	-0.001	0	0	-0	-0	-0	0	-0.008	-0.001	-0	0.009
Other	-0.001	-0	0	-0	-0	-0.001	-0	-0	0.001	-0	-0	-0	-0	-0	0	-0.007	-0.001	-0	-0
Male	0.001	0.014	0.001	0.007	0.004	0.011	-0.002	0.014	-0.008	0	-0.004	0.007	-0	0	-0.007	0.017	-0.016	-0	-0
Female	-0.001	-0.014	-0.001	-0.007	-0.004	-0.011	0.002	-0.014	0.008	-0	0.004	-0.007	0	-0	0.007	-0.017	0.016	0	0
<=50K	-0.025	-0.012	0	-0.026	-0.011	-0.025	-0.001	-0.003	0.015	-0.004	-0.003	-0.011	0	0	0.008	-0.013	0.011	-0.001	0.001
>50K	0.025	0.012	-0	0.026	0.011	0.025	0.001	0.003	-0.015	0.004	0.003	0.011	-0	-0	-0.008	0.013	-0.011	0.001	-0.001

The second output port gives spectral decomposition of PCA which contains eigenvalues and eigenvector. The eigenvalues gives the variation for each components. The PC1 always have maximum variation from the data. Eigenvector shows the loading which are weights given to each original dimension based on importance. Higher the value of loading indicates more variation in the component from that original variable. This output node become input node for **PCA apply Node.**

Spectral decomposition - 0:139 - PCA Compute (Components variations and)

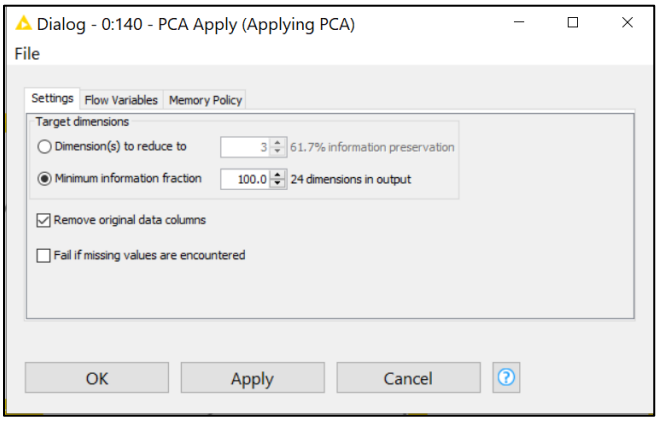
File Hilite Navigation View

Table "spectral decomposition" - Rows: 24 Spec - Columns: 25 Properties Flow Variables

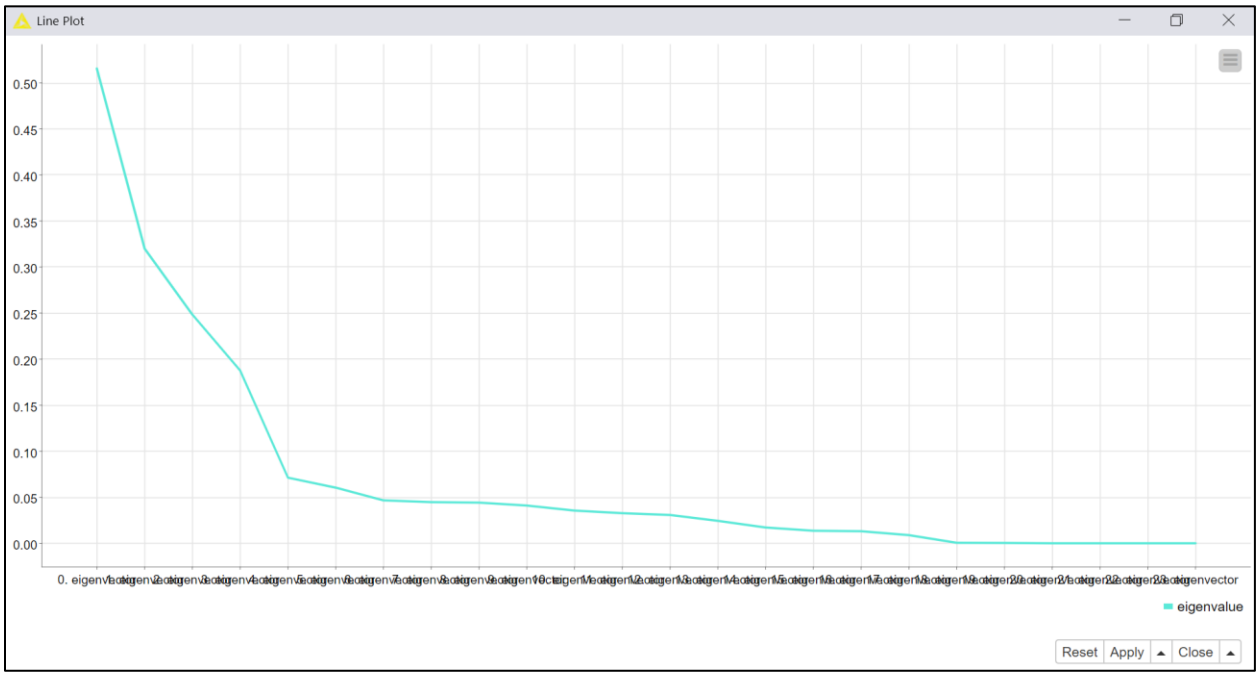
Row ID	eigenvalue	educati...	hours_...	t_fnlwgt	t_capit...	t_capit...	t_age	State-gov	Self-em...	Private	Federal...	Local-gov	Self-em...	Withou...	Never-...	missing	White	Black	Asian-P...
0. eigenvector	0.516	-0.047	-0.053	-0	-0.065	-0.029	-0.075	-0.002	-0.055	0.086	-0.01	-0.003	-0.041	0	0	0.026	-0.096	0.083	0.004
1. eigenvector	0.32	-0.099	-0.011	0.007	-0.09	-0.035	-0.085	-0.028	-0.011	0.182	-0.029	-0.062	-0.038	0	0	-0.014	-0.001	-0.001	-0.004
2. eigenvector	0.248	0.008	-0.007	-0.015	-0.017	-0.004	0.062	0.107	0.258	-0.886	0.072	0.193	0.077	0.001	0.001	0.177	-0.048	0.036	0.005
3. eigenvector	0.188	-0.01	0.002	0.027	0.008	0	0.009	0.005	-0.057	0.037	0.023	0.012	-0.013	-0	0	-0.008	-0.777	0.596	0.124
4. eigenvector	0.071	0.036	-0.043	0.021	-0.021	-0.004	-0.096	0.077	-0.817	-0.075	0.036	0.499	0.065	0	0	0.215	0.052	-0.013	-0.034
5. eigenvector	0.06	-0.081	-0.079	-0.001	0.019	-0.029	-0.091	0.084	-0.187	-0.056	0.045	-0.656	0.07	0	0	0.699	-0.021	-0.04	0.05
6. eigenvector	0.047	0.192	0.035	-0.064	-0.106	0.045	-0.184	0.507	-0.142	-0.077	0.105	-0.196	0.167	0.001	0	-0.366	-0.202	-0.369	0.484
7. eigenvector	0.045	-0.021	-0.044	-0.029	-0.366	0.179	-0.25	-0.438	0.176	0.052	-0.115	0.255	-0.218	-0.001	-0	0.289	-0.191	-0.328	0.422
8. eigenvector	0.044	0.007	0.02	-0.053	0.703	-0.317	0.312	-0.235	-0.018	0.042	-0.047	0.092	0.085	0.001	-0.001	0.082	-0.151	-0.266	0.341
9. eigenvector	0.041	0.058	-0.053	0.06	0.352	-0.329	-0.814	0.093	0.138	-0.033	-0.083	0.074	-0.186	-0.001	0.001	-0.005	0.033	0.07	-0.096
10. eigenvector	0.035	-0.002	0.079	0.018	-0.001	0.087	-0.294	-0.519	-0.059	-0.077	0.214	-0.115	0.726	0	0.001	-0.172	0.011	0.043	-0.046
11. eigenvector	0.033	-0.13	-0.008	-0.014	-0.448	-0.849	0.105	-0.077	-0.024	-0.019	0.171	-0.014	0.024	0	0	-0.061	-0.006	-0.03	0.022
12. eigenvector	0.031	0.073	-0.024	-0.029	0.117	0.109	-0.017	-0.174	-0.068	-0.037	0.856	-0.073	-0.43	0.001	0	-0.074	0.015	-0.022	-0.002
13. eigenvector	0.024	0.944	0.12	-0.044	-0.094	-0.116	0.066	-0.111	0.03	0.048	-0.062	-0.037	-0.005	-0.001	-0.002	0.141	0.04	0.098	-0.04
14. eigenvector	0.017	0.041	-0.075	0.947	-0.005	-0.004	0.059	-0.003	0.011	-0.012	0.028	-0.011	-0	0	0	-0.013	0.064	0.019	0.187
15. eigenvector	0.014	-0.066	0.85	0.202	-0.007	-0.006	-0.03	0.031	-0.025	-0.005	0.005	-0.005	-0.074	-0.002	-0.001	0.076	-0.133	-0.164	-0.208
16. eigenvector	0.013	0.138	-0.482	0.194	-0.008	0.004	0.05	-0.014	0.019	0.004	-0.007	-0.012	0.051	0	0	-0.042	-0.234	-0.272	-0.397
17. eigenvector	0.009	-0.01	0.01	0.095	-0.003	0.004	-0.006	-0.003	0	0.002	-0.004	-0.006	0.006	-0	0	0.003	0.042	0.041	0.062
18. eigenvector	0	-0.001	-0.003	-0	0.001	-0	0	0.147	0.146	0.145	0.148	0.146	0.147	-0.915	-0.11	0.146	0	0	0.001
19. eigenvector	0	-0.001	-0.001	0.001	-0	-0	-0.002	0.102	0.101	0.101	0.102	0.101	0.102	0.226	-0.936	0.101	0	0	-0
20. eigenvector	0	-0	-0	-0	0	0	0	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	-0.333	0.005	0.005	0.005
21. eigenvector	0	0	0	-0	0	0	-0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.003	0.003
22. eigenvector	-0	0	-0	0	-0	-0	0	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.004	-0.024	-0.024	-0.024
23. eigenvector	-0	0	0	-0	-0	0	-0	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.447	0.447	0.447

The third port shows the transformation model which then becomes input for PCA apply Node.

Then PCA apply Node has been used to generate PC and giving the following configuration as retaining 100% information which means same number of PC has been generated as original dimensions.

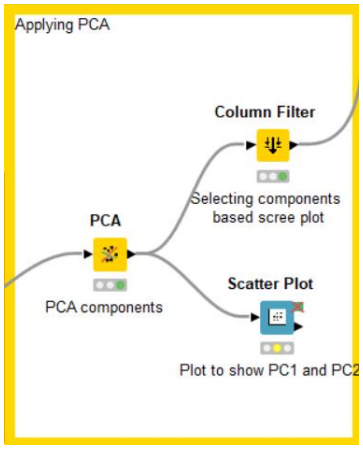


After generating all principle components, variation has been explored for all the PC to select number of PC for final step. The screen plot has been used explain variation in each PC. X-axis is the number of PC which is 24 whereas, y-axis is the percentage of the variation explained by a component. Usually the PC is selected on first tail or elbow observed in the line which is here PC=2 less than 30% variation explained. After PC=5 there is less than 5% variation explain in the data.

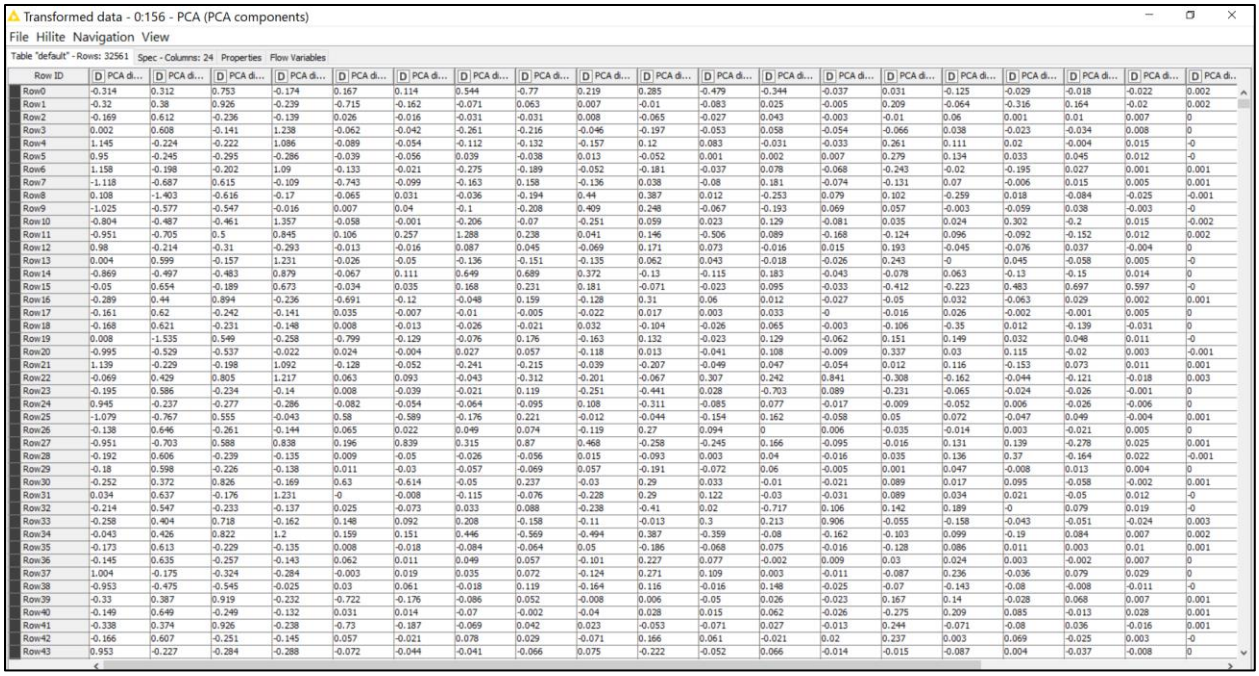


2. Applying PCA:

The following Knime workflow is applying PCA using **PCA node** and then selecting the number of PC as explained above. The **scatterplot** is showed between first two components which depicts pattern as most of the variation of the data is explained by the first components.

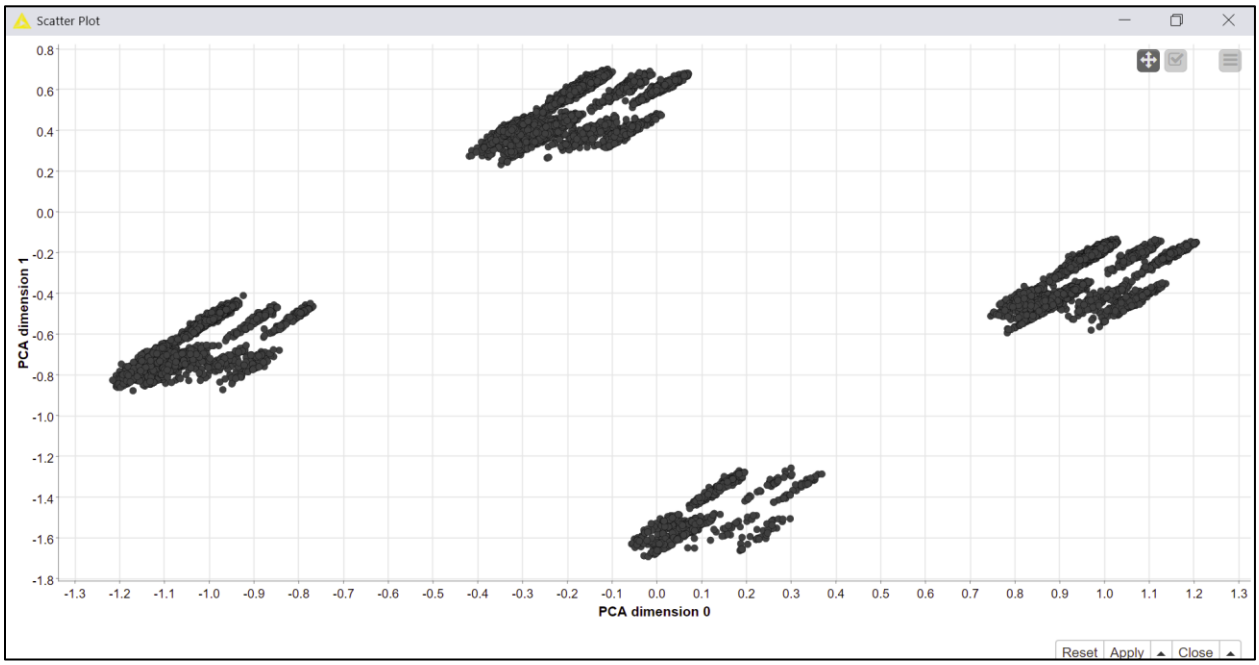


After applying PCA, 24 PC has been showed in the screen below.

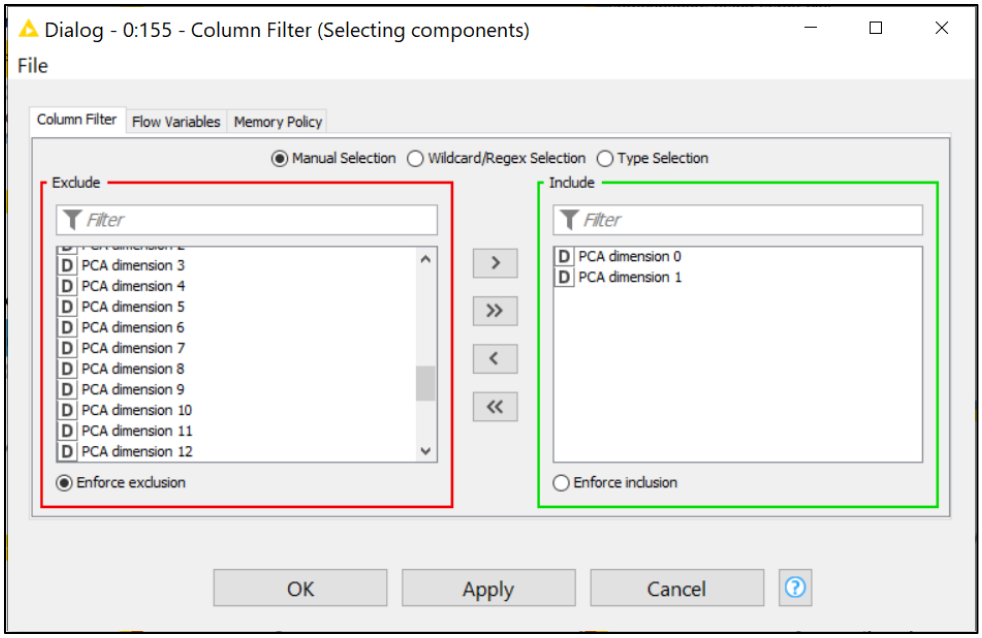


Row ID	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...	D PCA di...
Row0	-0.314	0.312	0.753	-0.174	0.167	0.114	0.544	-0.77	0.219	0.285	-0.479	-0.344	-0.037	0.031	-0.125	-0.029	-0.018	-0.022	0.002
Row1	-0.32	0.38	0.926	-0.239	-0.715	-0.162	-0.071	0.063	0.007	-0.01	-0.083	0.025	-0.005	0.209	-0.064	-0.316	0.164	-0.02	0.002
Row2	-0.169	0.612	-0.236	-0.139	0.026	-0.016	-0.031	-0.031	0.008	-0.065	-0.027	0.043	-0.003	-0.01	0.06	0.001	0.01	0.007	0
Row3	0.002	0.608	-0.141	1.238	-0.062	-0.042	-0.261	-0.216	-0.046	-0.197	-0.053	0.058	-0.054	-0.066	0.038	-0.023	-0.034	0.008	0
Row4	1.145	-0.224	-0.222	1.086	-0.089	-0.054	-0.112	-0.132	-0.157	0.12	0.083	-0.031	-0.033	0.261	0.111	0.02	-0.004	0.015	-0
Row5	0.95	-0.245	-0.295	-0.286	-0.039	-0.056	0.039	-0.038	0.013	-0.052	0.001	0.002	0.007	0.279	0.134	0.033	0.045	0.012	-0
Row6	1.158	-0.198	-0.202	1.09	-0.133	-0.021	-0.275	-0.189	-0.052	-0.181	-0.037	0.078	-0.068	-0.243	-0.02	-0.195	0.027	0.001	0.001
Row7	-1.118	-0.687	0.615	-0.109	-0.743	-0.099	-0.163	0.158	-0.136	0.038	-0.08	0.181	-0.074	-0.131	0.07	-0.006	0.015	0.005	0.001
Row8	0.108	-1.403	-0.616	-0.17	-0.065	0.031	-0.036	-0.194	0.44	0.387	0.012	-0.253	0.079	0.102	-0.259	0.018	-0.084	-0.025	-0.001
Row9	-1.025	-0.577	-0.547	-0.016	0.007	0.04	-0.1	-0.208	0.409	0.248	-0.067	-0.193	0.069	0.057	-0.003	-0.059	0.038	-0.003	-0
Row10	-0.804	-0.487	-0.461	1.357	-0.058	-0.001	-0.206	-0.07	-0.251	0.059	0.023	0.129	-0.081	0.035	0.024	0.302	-0.2	0.015	-0.002
Row11	-0.951	-0.705	0.5	0.845	0.106	0.257	1.288	0.238	0.041	0.146	-0.506	0.089	-0.168	-0.124	0.096	-0.092	-0.152	0.012	0.002
Row12	0.98	-0.214	-0.31	-0.293	-0.013	-0.016	0.087	0.045	-0.069	0.171	0.073	-0.016	0.015	0.193	-0.045	-0.076	0.037	-0.004	0
Row13	0.004	0.599	-0.157	1.231	-0.026	-0.05	-0.136	-0.151	-0.135	0.062	0.043	-0.018	-0.026	0.243	-0	0.045	-0.058	0.005	-0
Row14	-0.869	-0.497	-0.483	0.879	-0.067	0.111	0.649	0.689	0.372	-0.13	-0.115	0.183	-0.043	-0.078	0.063	-0.13	-0.15	0.014	0
Row15	-0.05	0.654	-0.189	0.673	-0.034	0.035	0.168	0.231	0.181	-0.071	-0.023	0.095	-0.033	-0.412	-0.223	0.483	0.697	0.597	-0
Row16	-0.289	0.44	0.894	-0.236	-0.691	-0.12	-0.048	0.159	-0.128	0.31	0.06	0.012	-0.027	-0.05	0.032	-0.063	0.029	0.002	0.001
Row17	-0.161	0.62	-0.242	-0.141	0.035	-0.007	-0.01	-0.005	-0.022	0.017	0.003	0.033	-0	-0.016	0.026	-0.002	-0.001	0.005	0
Row18	-0.168	0.621	-0.231	-0.148	0.008	-0.013	-0.026	-0.021	0.032	-0.104	-0.026	0.065	-0.003	-0.106	-0.35	0.012	-0.139	-0.031	0
Row19	0.008	-1.535	0.549	-0.258	-0.799	-0.129	-0.076	0.176	-0.163	0.132	-0.023	0.129	-0.062	0.151	0.149	0.032	0.048	0.011	-0
Row20	-0.995	-0.529	-0.537	-0.022	0.024	-0.004	0.027	0.057	-0.118	0.013	-0.041	0.108	-0.009	0.337	0.03	0.115	-0.02	0.003	-0.001
Row21	1.139	-0.229	-0.198	1.092	-0.128	-0.052	-0.241	-0.215	-0.039	-0.207	-0.049	0.047	-0.054	0.012	0.116	-0.153	0.073	0.011	0.001
Row22	-0.069	0.429	0.805	1.217	0.063	0.093	-0.043	-0.312	-0.201	-0.067	0.307	0.242	0.841	-0.308	-0.162	-0.044	-0.121	-0.018	0.003
Row23	-0.195	0.586	-0.234	-0.14	0.008	-0.039	-0.021	0.119	-0.251	-0.441	0.028	-0.703	0.089	-0.231	-0.065	-0.024	-0.026	-0.001	0
Row24	0.945	-0.237	-0.277	-0.286	-0.082	-0.054	-0.064	-0.095	0.108	-0.311	-0.085	0.077	-0.017	-0.009	-0.052	0.006	-0.026	-0.006	0
Row25	-1.079	-0.767	0.555	-0.043	0.58	-0.589	-0.176	0.221	-0.012	-0.044	-0.154	0.162	-0.058	0.05	0.072	-0.047	0.049	-0.004	0.001
Row26	-0.138	0.646	-0.261	-0.144	0.065	0.022	0.049	0.074	-0.119	0.27	0.094	0	0.006	-0.035	-0.014	0.003	-0.021	0.005	0
Row27	-0.951	-0.703	0.588	0.838	0.196	0.839	0.315	0.87	0.468	-0.258	-0.245	0.166	-0.095	-0.016	0.131	0.139	-0.278	0.025	0.001
Row28	-0.192	0.606	-0.239	-0.135	0.009	-0.05	-0.026	-0.056	0.015	-0.093	0.003	0.04	-0.016	0.035	0.136	0.37	-0.164	0.022	-0.001
Row29	-0.18	0.598	-0.226	-0.138	0.011	-0.03	-0.057	-0.069	0.057	-0.191	-0.072	0.06	-0.005	0.001	0.047	-0.008	0.013	0.004	0
Row30	-0.252	0.372	0.826	-0.169	0.63	-0.614	-0.05	0.237	-0.03	0.29	0.033	-0.01	-0.021	0.089	0.017	0.095	-0.058	-0.002	0.001
Row31	0.034	0.637	-0.176	1.231	-0	-0.008	-0.115	-0.076	-0.228	0.29	0.122	-0.03	-0.031	0.089	0.034	0.021	-0.05	0.012	-0
Row32	-0.214	0.547	-0.233	-0.137	0.025	-0.073	0.033	0.088	-0.238	-0.41	0.02	-0.717	0.106	0.142	0.189	-0	0.079	0.019	-0
Row33	-0.258	0.404	0.718	-0.162	0.148	0.092	0.208	-0.158	-0.11	-0.013	0.3	0.213	0.906	-0.055	-0.158	-0.043	-0.051	-0.024	0.003
Row34	-0.043	0.426	0.822	1.2	0.159	0.151	0.446	-0.569	-0.494	0.387	-0.359	-0.08	-0.162	-0.103	0.099	-0.19	0.084	0.007	0.002
Row35	-0.173	0.613	-0.229	-0.135	0.008	-0.018	-0.084	-0.064	0.05	-0.186	-0.068	0.075	-0.016	-0.128	0.086	0.011	0.003	0.01	0.001
Row36	-0.145	0.635	-0.257	-0.143	0.062	0.011	0.049	0.057	-0.101	0.227	0.077	-0.002	0.009	0.03	0.024	0.003	-0.002	0.007	0
Row37	1.004	-0.175	-0.324	-0.284	-0.003	0.019	0.035	0.072	-0.124	0.271	0.109	0.003	-0.011	-0.087	0.236	-0.036	0.079	0.029	0
Row38	-0.953	-0.475	-0.545	-0.025	0.03	0.061	-0.018	0.119	-0.164	0.116	-0.016	0.148	-0.025	-0.07	-0.143	-0.08	-0.008	-0.011	-0
Row39	-0.33	0.387	0.919	-0.232	-0.722	-0.176	-0.086	0.052	-0.008	0.006	-0.05	0.026	-0.023	0.167	0.14	-0.028	0.068	0.007	0.001
Row40	-0.149	0.649	-0.249	-0.132	0.031	0.014	-0.07	-0.002	-0.04	0.028	0.015	0.062	-0.026	-0.275	0.209	0.085	-0.013	0.028	0.001
Row41	-0.338	0.374	0.926	-0.238	-0.73	-0.187	-0.069	0.042	0.023	-0.053	-0.071	0.027	-0.013	0.244	-0.071	-0.08	0.036	-0.016	0.001
Row42	-0.166	0.607	-0.251	-0.145	0.057	-0.021	0.078	0.029	-0.071	0.166	0.061	-0.021	0.02	0.237	0.003	0.069	-0.025	0.003	-0
Row43	0.953	-0.227	-0.284	-0.288	-0.072	-0.044	-0.041	-0.066	0.075	-0.222	-0.052	0.066	-0.014	-0.015	-0.087	0.004	-0.037	-0.008	0

Scatterplot between first two components shows strong grouping because these two components contain variation of the data as can be seen below.



Column filter Node filters two first components in the following screen.



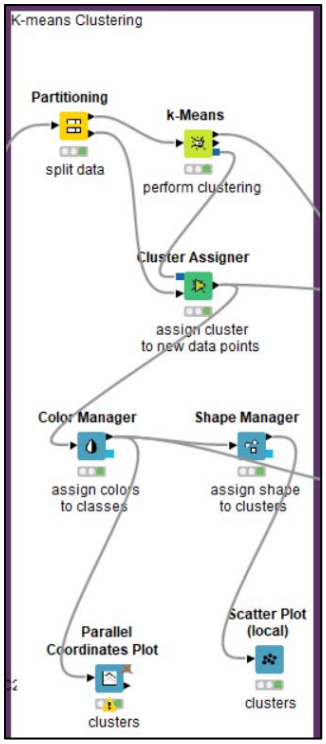
Description/ Interpretation of PC:

Principal Component	Variables
PC1 & PC2	Workclass, Race, sex, Salary

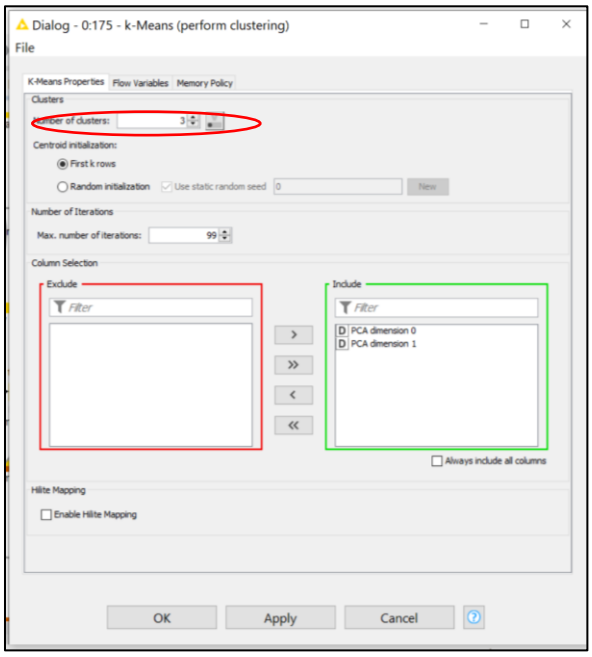
K-Means Clustering:

K-means clustering is partition-based clustering which generated clustered based on k. k is specified in the algorithm before applying clustering. There is always difficult to decide what should be the k value. But it can decided by applying the clustering again and again to check which number of k gives best clustering. K-means has been selected as all the variables have been converted to numeric and distance can be calculated using k-means. Second k-means is good in grouping the observation as compared to agglomerative method. However, this algorithm is sensitive to noise and outlier. Smoothing and normal distribution has been applied to overcome this problem.

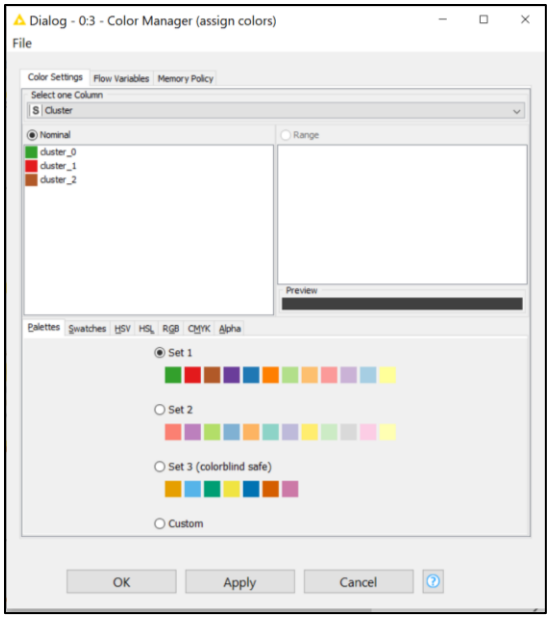
The following screen shows **Partitioning Node** which has been used to split data into two 80-20 partition to run clustering on first partition.



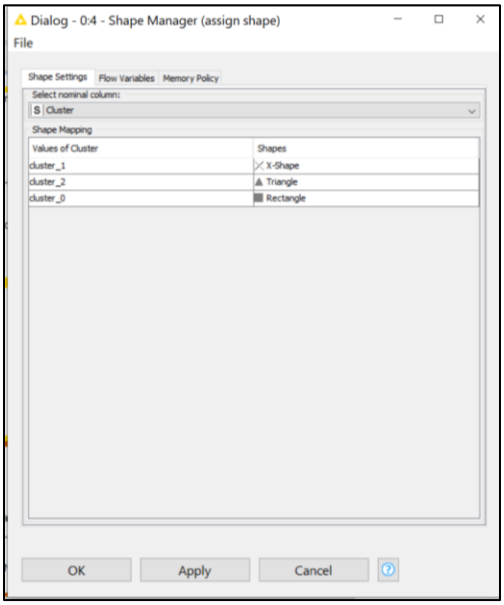
Then clustering model to be used to assign 20% data into clusters using **Cluster Assigner Node**. In k-means Node clusters=3 has been set.



Next node is **Color Manager,** it has been used to assign color to different clusters.



The figure below shows **Shape Manager Node** which is used to assign different shapes to represent different clusters.



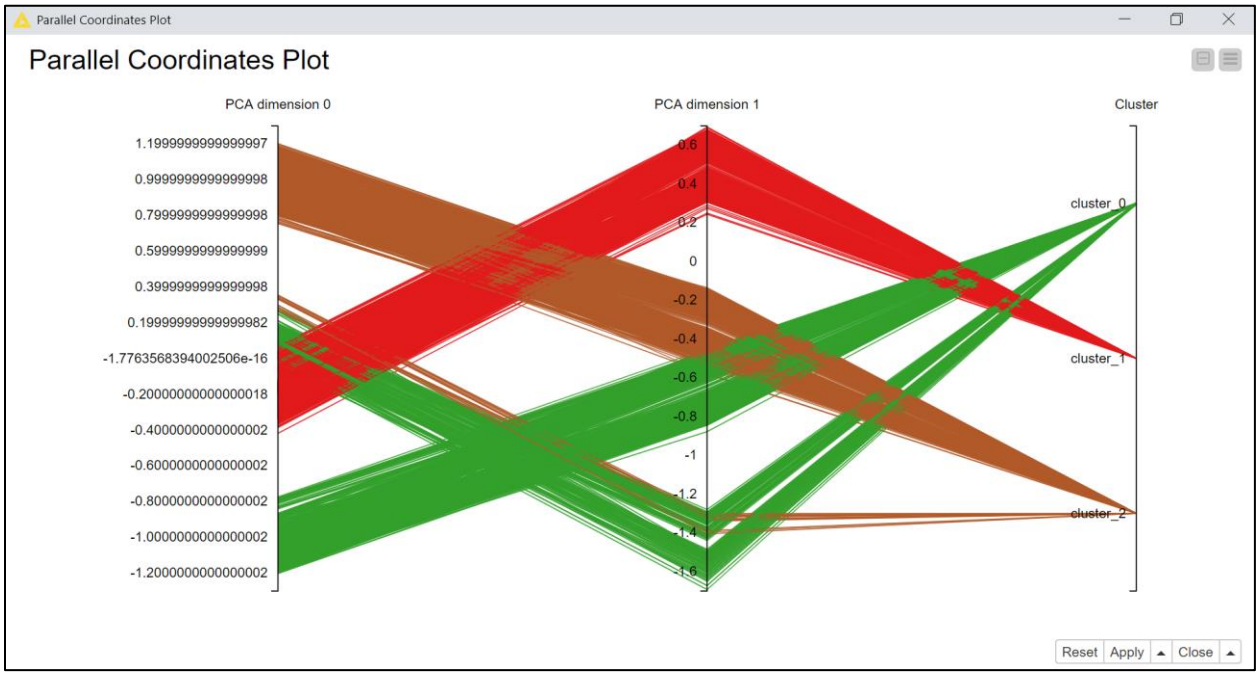
Parallel Coordinates Node shows three cluster with three color and two PC mapping to clusters.

Parallel coordinates explain how three clusters values varies.

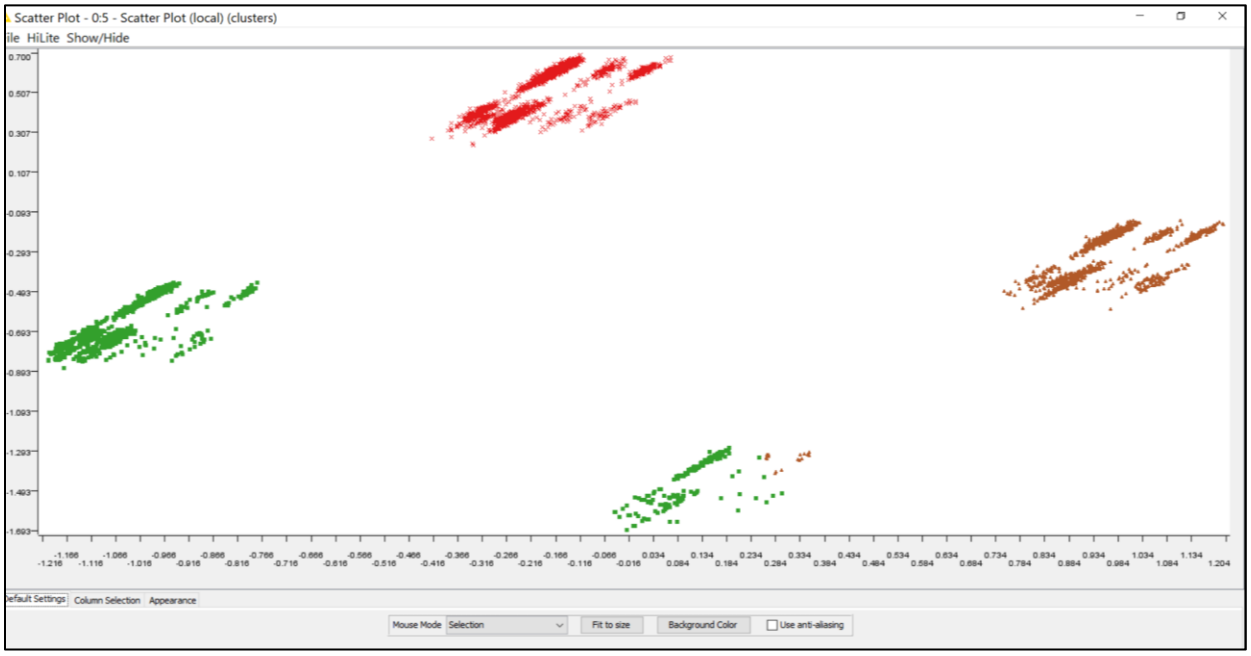
Cluster 0 which is weakly grouped have some values in the middle and some values very low for PC0 and same for PC1.

Cluster 1, have value for PC0 in the middle and then PC1 it has high value.

Cluster 2, PC0 have high value and PC1 have value in the middle.



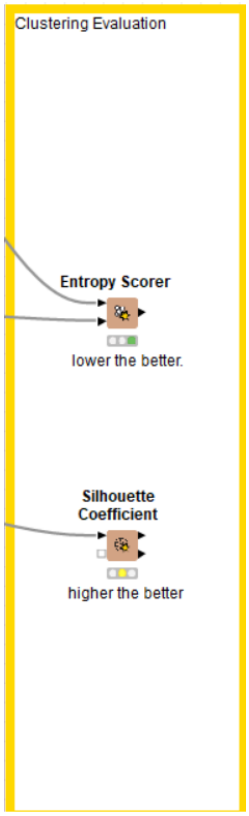
Scatter Plot Node is also used to show data in clusters.



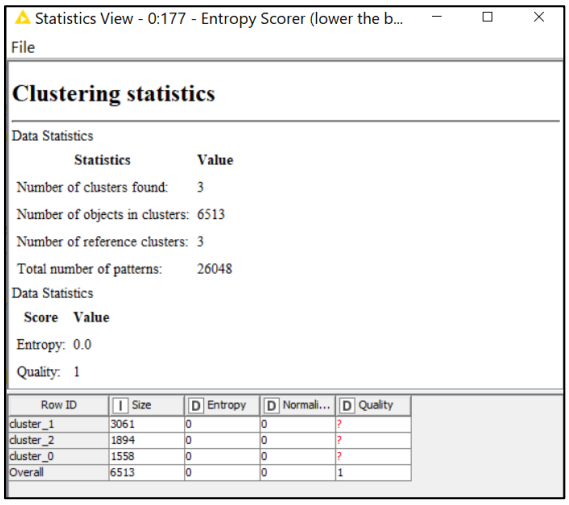
Clustering Evaluation:

1. **Entropy Scorer Node:**

To evaluate the results of clustering, **Entropy Scorer Node** has been used.



Entropy Scorer calculates a score given a reference clustering from **K-Means Node** which contains **clustering** column in one input and then the table where the clusters have been assigned using **Cluster Assigner Node**. It compares both columns and gives entropy score. The smaller the value of entropy, the better is the clustering. Cluster Quality value is also used and the best possible value for quality is one. The following screen shows entropy=0 which is best possible value indicating the clustering results good. Quality is 1 which also indicates best.

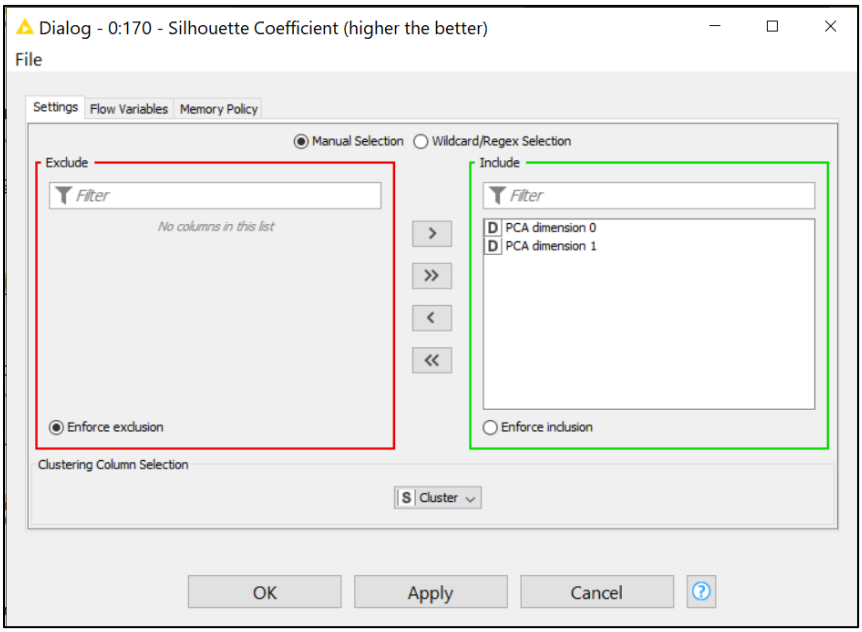


2. Silhouette Coefficient

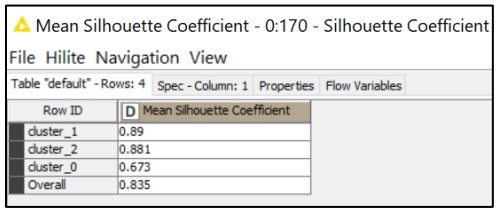
Silhouette coefficient is used to evaluate clustering performance. It uses the following formula

(b-a)/max(a,b)

a is the mean value of the distance of intra-cluster whereas b is the mean value of distance of inter-cluster to the closest cluster[3]. The value range from -1.0-1. The higher the value is the better clustering. It takes a table as input which should have input columns and a column of cluster to calculate the distance and score of silhouette coefficient.



The following screen shows calculated coefficient for the three clusters. Cluster_0 has low score as compared to others. Overall score is =0.835 which is very close to zero which means clustering is defined well.



[3] https://nodepit.com/node/org.knime.base.node.mine.cluster.eval.silhouette.SilhouetteCoefficientNodeFactory

References

Chet Lemon, C. Z. K. M., n.d. Predicting if income exceeds $50,000 per year based on 1994 US Census Data with.

Glenn J. Myatt, W. P. J., 2009. *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications.* s.l.:John Wiley & Sons.

Larose, D. T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining.* s.l.:John Wiley & Sons.

Myatt, G. J., 2007. *Making Sense of Data: A practical Guide to Exploratory data Analysis and Data Mining.* New Jersey: John Wiley & Sons.

Nishida, K., 2018. *Visualizing K-Means Clustering Results to Understand the Clusters Better,* s.l.: Medium.

Ray, S., 2016. *A comprehensive Guide to Data Exploration,* s.l.: Analytics Vidhya.

Zhu, H., 2016. Predicting Earning Potential using the Adult Dataset.