A close up of a flower

Description automatically generated

**Using Data Mining Techniques to interpret if income exceeds $50,000 per year based on 1994 US Census Data with**

**Presented by-**

Nirvi Khiyani [khyiyanin@lsbu.ac.uk]

Student Id: 3901390

Module Code: CSI\_7\_DMA\_1920

MSc Data Science

**Group- I**

**Guided By** -**Dr D Chen**

**Table of Contents**

1. Abstract 3
2. Introduction 3
3. Business Understanding 3

3.1 Data Mining Objective 3

1. Data Understanding 4

4.1 Describe the Data 4

1. Data Preparation 5
   1. Data Pre-processing 5
   2. Exploring the data 5
   3. Removal of Features 7
2. Modelling 7

6.1 Predictive Modelling 8

6.2 Descriptive Modelling 12

7. Evaluation 16

7.1 Evaluation of Results for Predictive Modelling 16

7 .2 Evaluation of Results for Descriptive Modelling 17

8. Conclusion 22

9. References 22

1. **Abstract:**

In this report we have implemented Data Mining methodology which is CRISP model that helps to identify the problem that individual whose salary exceeds than $50, 000 with the given demographic attributes like age, education and its employment type. This model starts from business understanding till evaluation and deployment. Concepts such as statistical application analysis which suggest the usefulness of attributes, outlier detection, missing values feature reduction, data bias and data transformations which are discussed in the below sections. The business objective is to use predictive capabilities of the models proposed to predict the income of the individuals. For this coursework of Data Mining, we have examined the Census Income dataset available, by using Data Mining techniques like Decision tree, Logistic regression and Cluster Analysis .This is the report of the tasks performed to predict individual’s income is exceeds $50,000 per in the year 1994 depending on the various variables of the census data.

1. **Introduction :**

In our project, we have investigated data mining of demographic data which is achieved by making use of classification models that are capable of identifying individuals who have salary more than the given value accurately. The source of the used data is Census bureau database and hence, stated as Adult-dataset and has attributes like current employment, education level and age. The classification model are used to select the candidates who are offered a new service which is offered by the project targeting sponsor whose salary exceeds $50, 000. Each attribute’s predictive significance is described and transformations are applied .This report states the tasks of the iterative process which includes data-preparation, its modelling and evaluating the data that has data formatting as well, quality issues or consistency issue, removing or rejecting the non-significant attributes and removing of outliers or missing values in the data.

In this analysis, we have used CRISP methodology of Data mining, which has various phases starting from Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation. Also, in this we have made use of SAS software to develop and evaluate different Descriptive and Predictive models that helps to predict which model is best as per Adult dataset.

1. **Business Understanding:**

The Adult dataset was extrcated from Census database having 32561 records.

Our project goal is to create the models that accurately identify the individuals in the Adult dataset who have income more than $50, 000.

**3.1 Data mining objective:**

Data Mining objective is to create and implement Descriptive and Predictive models that accurately identifies the individuals in the Adult dataset who have income more than $50, 000 by using attributes like age, education, Marital status, Relationship and Income. For each of the used models, the return-on-investment and the details of all the data transformations carried out is given along with useful inputs like the most significant attributes.

1. **Data Understanding :**

**4.1 Describe the data:**

The repository of Census dataset consists of 32561 entries extracted from the Census database and a binomial label which indicates the income is greater than or less than $50, 000 and is referred as <50 or >=50.The attributes in the original data are listed below:

● Age: the age of individual is greater than 17 and not greater than 90.

● Workclass: Term that represents individual’s employment status which are Local-gov, State-gov, Without-pay, Never-worked, Private, Self-emp-not-inc, Self-emp-inc, Federal-gov,

●Fnlwgt: In simpler terms, this is the number of people the census believes entry represents

Which is greater than 0 and less than or equal to 1184622?

● education : indicates the higher level of education an individual, which has the values for education attribute like Preschool, 1st4th, , 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Bachelors, Some-college, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, Masters and Doctorate

● education-num: The education level is the numeric attribute greater than 0 and less than or equal to 16

● Marital-status: This attribute indicates the individual’s marital status which includes values like Married-civ-spouse, Married-AF-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent.

● Occupation: Individuals occupation is like Machine-op-inspct, Adm-clerical, Tech-support, Craft-repair, other-service, Sales, Exec- managerial, Prof-specialty, Handlers-cleaners,

Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv and Armed-Forces.

● Relationship: This attribute denotes that individual is relative to others. For example an

Individual can be a Husband. Each entry only has one relationship attribute and is

Somewhat redundant with marital status. The values are Husband, Not-in-family, other-relative, Wife, Own-child, and Unmarried.

● Race: Race of an individual’s is White, Asian-Pac-Islander, Amer-Indian-Eskimo, Black and Other.

● Sex: individual’s sex which can be either female or male.

● Capital-gain: It is individual capital-gain which is integer greater than or equal to 0

● Capital-loss: It is individual capital-loss which is integer greater than or equal to 0

● Hours-per-week: Individual’s hours per week which is integer greater than or equal to 0

● Native-country: Origin of an Individual which is United-States, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Portugal,

Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia,

, Thailand, Yugoslavia, El-Salvador,Trinadad & Tobago, Peru, Hong, Holand-Netherlands, Hungary, Guatemala, Nicaragua, Scotland.

● Income (the label): Indicates the individual income which is either less than 50K or greater than 50K

1. **Data Preparation:**
   1. **Data Pre-processing:**

Data Pre-processing is the transformation technique that requires transforming the data in the format that is acceptable to target application. In our dataset, we have transformed data into SAS acceptable as we have to develop and evaluate the model in SAS application. It includes various transformations:

1. We have updated the ? (Missing values) by full stop.
2. We have replaces – by \_ , as SAS does not accept – neither in data nor in attribute name.
3. We have converted Income in Binary (because income is having only two values which is >=50 and <50).
4. Also, for Native-Country attribute, most of the individuals are from United-States and hence, entries other than United\_States are replaced by Non\_United\_States.

After doing the steps from step1 to 4, we have imported the data in SAS Enterprise Guide and then from Enterprise Guide we have exported it to Enterprise Miner to analyse the data.

* 1. **Exploring the data:**

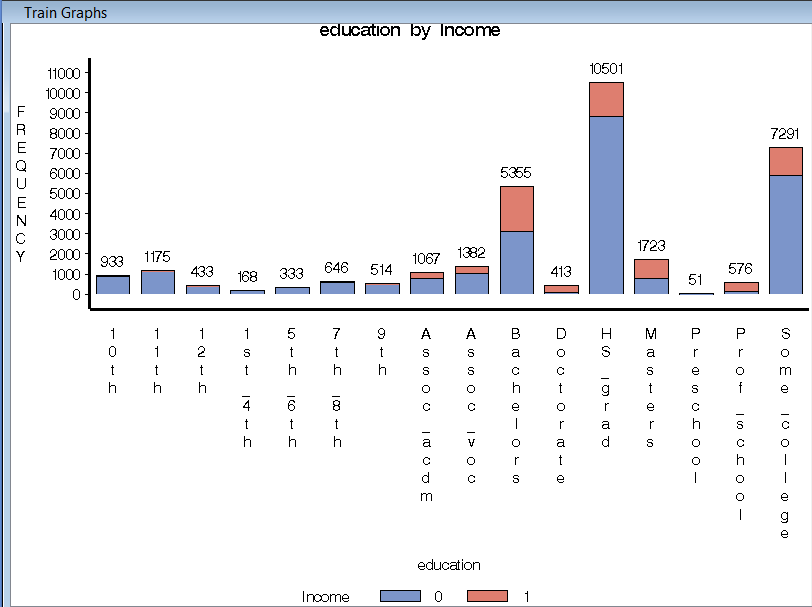
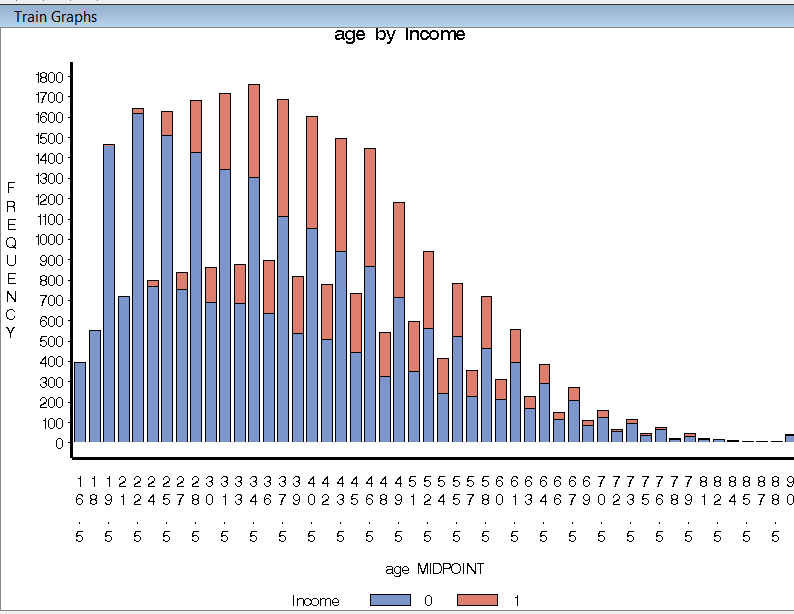
There are several attributes which are significant and are mandatory to predict the individual’s income and they are age, Marital Status, Relationship, Sex etc. Hence, there exists the relation between the above mentioned attributes which can be stated from the below graphs taken from SAS:

Graph 1 states the relationship between Age and Income and it can be predicted that Individuals in the age range 30 to 70 are having income greater than $50, 000 and is less than $50, 000 once individual is more than 70 years of age.

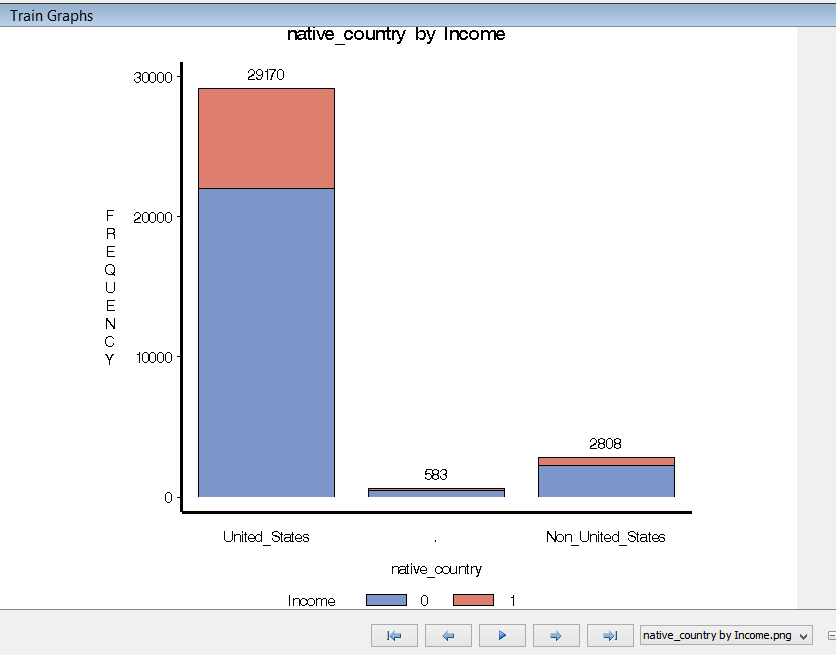
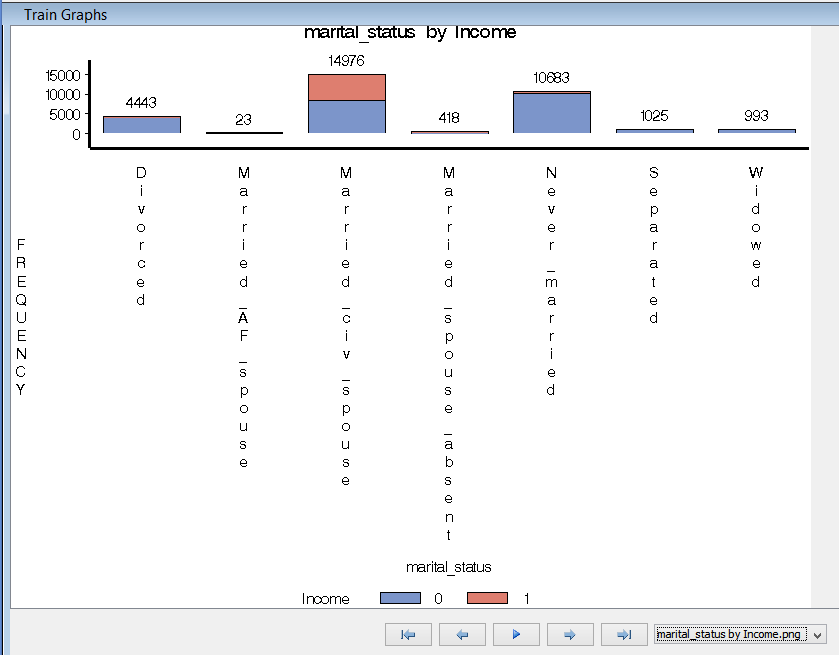
Graph 2 states the relation between education and income and it is predicted that till 9th Std, the income of an individual is less than $50, 000 and for Hs-grade degree the count of earning more than $50, 000 is around 1500 individuals.

Graph 3: It is about marital status and Income and it can be predicted that mostly individuals who are married-civil-partner, in which married-civil-partner is having more than 4500 individuals whose income exceeds than $50, 000.

Graph 4: Native country is not the significant attributes, but it bifurgating the data in two- categories and it is predicted that more than 7000 individuals who are from united states have income greater than $50, 0000, however the individual number in non-united-states is less than 500 individuals who earn more than $50, 000.



Graph 1 Graph 2



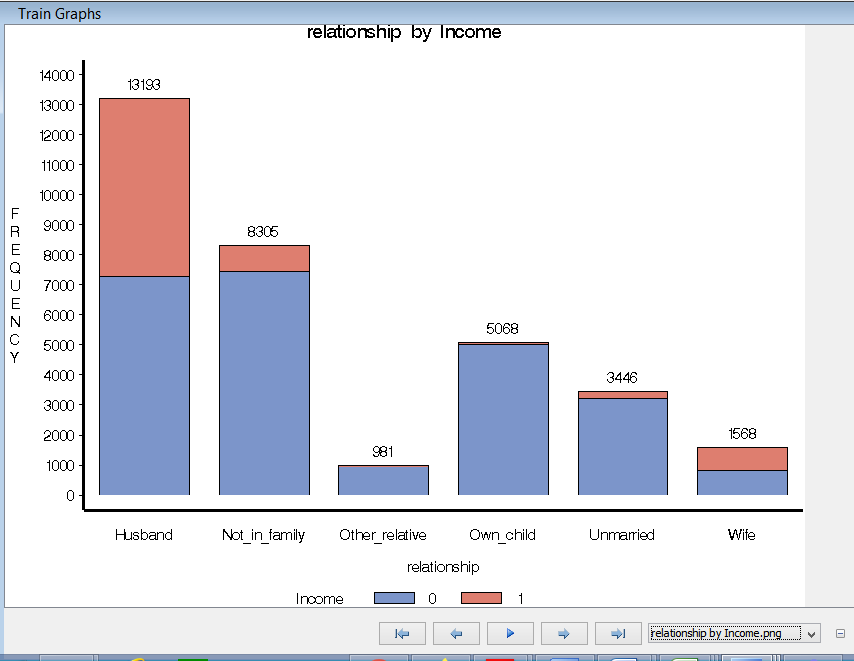
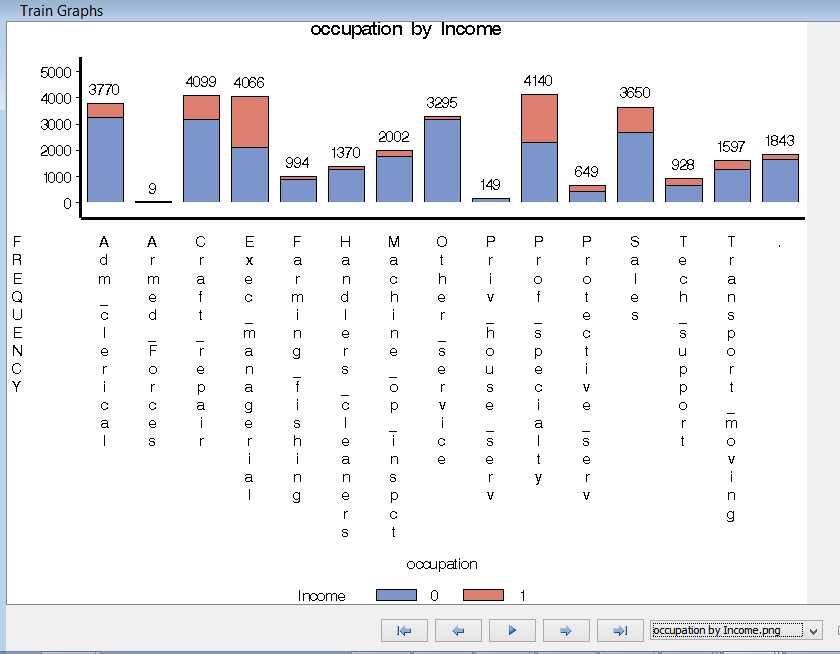
Graph 3 Graph 4

Graph 5: Represents the relation of Occupation with Income and it is suspected that individuals who earn more than $50, 000 are either Prof-Speciality, Exec-Manager or in Sales.

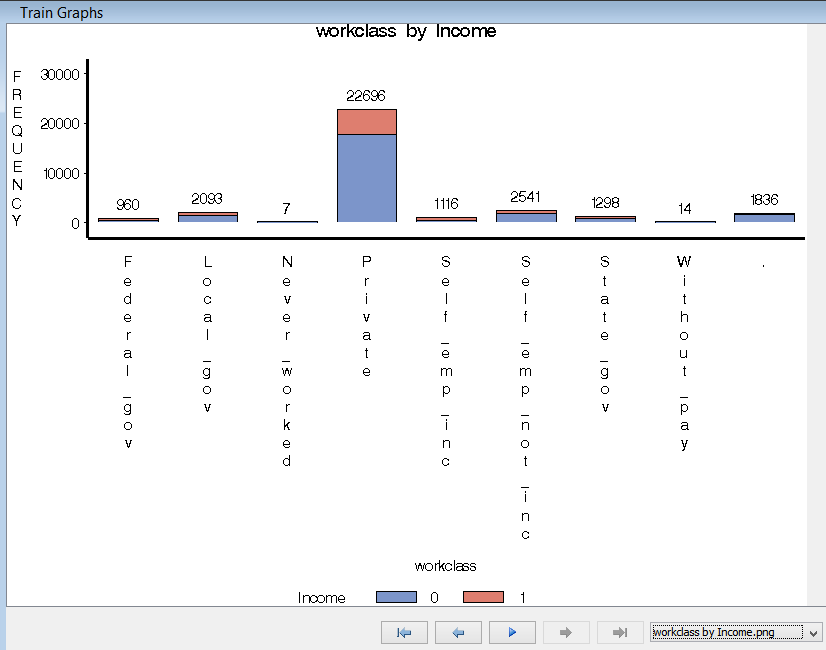
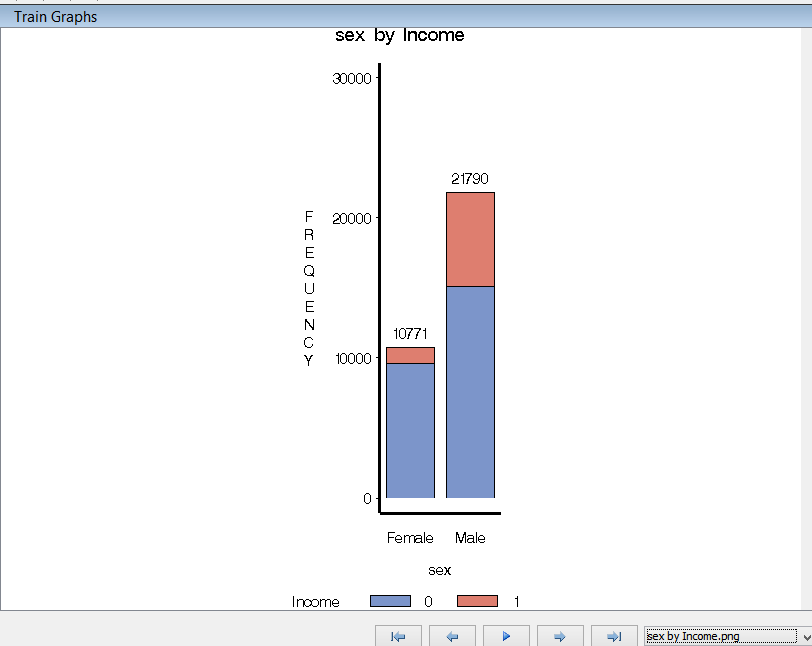
Graph 6: represent the association between Relationship and Income where Relationship is the most significant attribute and hence, it can be predicted that individuals who earn more than $50, 000 are either are husband, wife, not-in-family.

Graph 7: It is about sex and income and it can be said that count of males is more than 7000 whose income exceeds $50,000.

Graph 8: It is about workclass and income and it is clear that individuals who are in Private sector earn more than $50, 000 when compared to any other workclass.



Graph 5 Graph 6



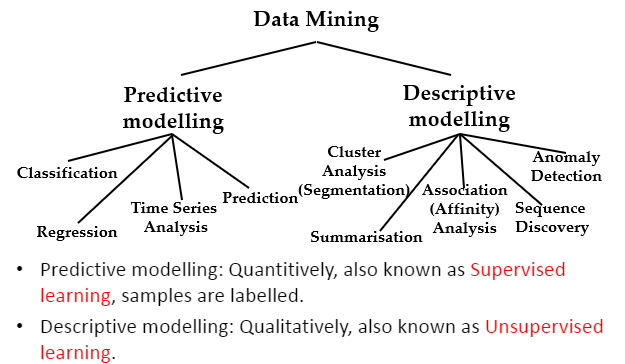
Graph 7 Graph 8

**5.3 Removal of Features:**

We have opted to not use the attributes like ‘fnlwgt’, ‘capital-gain/loss’, education-num, race, hours-per-week and hence we have rejected those attributes. Because, either these attributes were having bad data or it is not useful to our analysis**.**

1. **Modelling:**

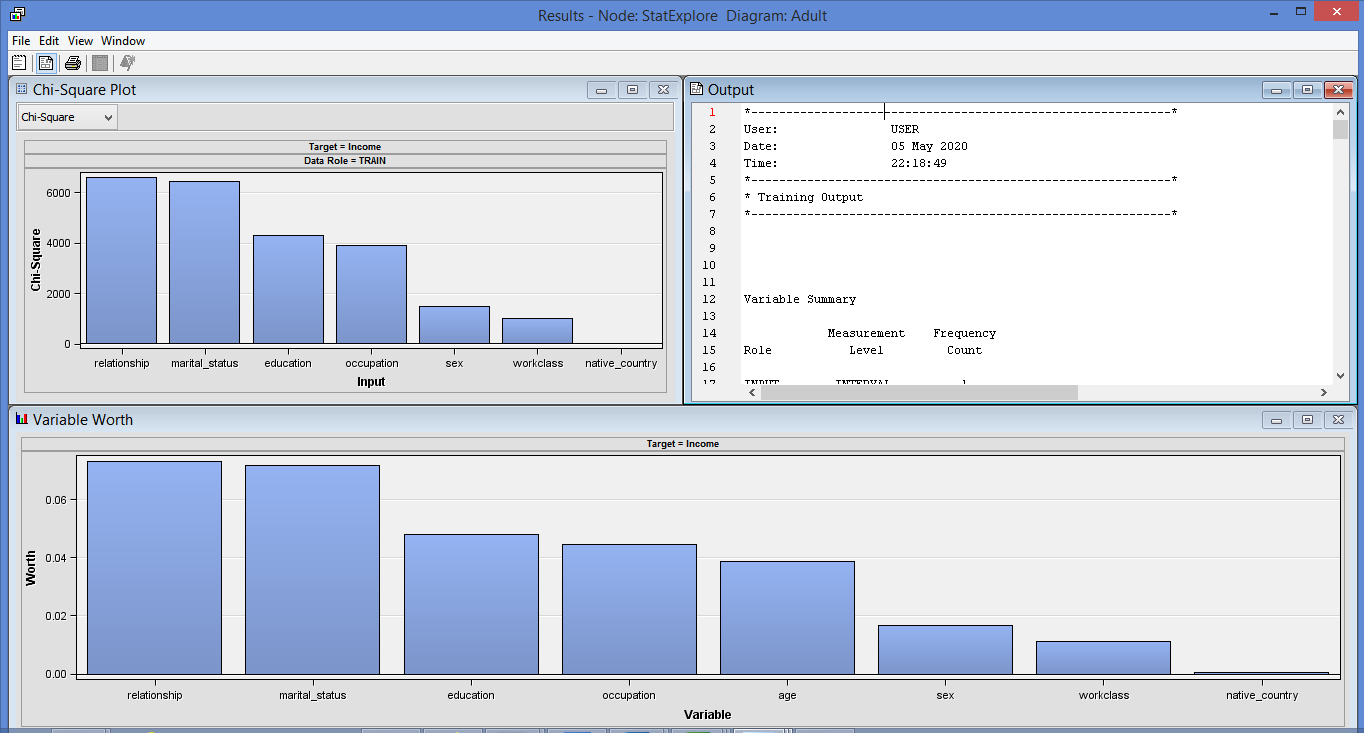
In Data Mining, there are Predictive and Descriptive models which can be further divided as stated under:



**6.1 Predictive Modelling:**

The original dataset contains a distribution of around 24% entries that are labeled with >50k and remaining 76% entries are labelled with <=50k. For this, dataset is divided into Validation and Training sets by using Data Partition node in SAS that will maintain the above distribution. All above graphs and statistics pertain to the training set.

|  |  |  |
| --- | --- | --- |
| Income | Number | Percentage |
| <=50 | 24720 | 75.91% |
| >50 | 7840 | 24.07% |

The below mentioned window is StatExplorer node window which states that relationship, marital-status and education are the most significant attribute, however, native-country and workclass are least significant attribute. 

Graph for Co-relation between the required attributes

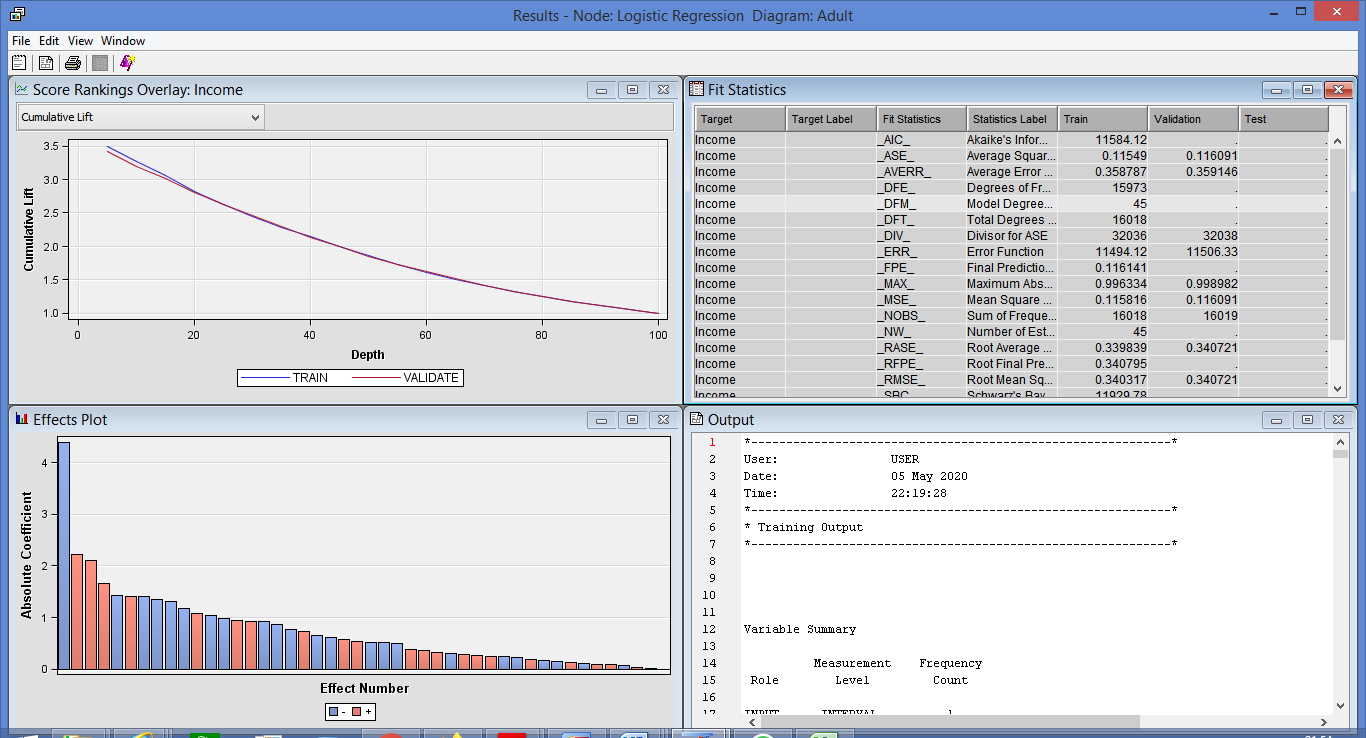
After the co-relation, the data is portioned into two sets, Training and Validation with both 50% as in the below table:

|  |  |  |
| --- | --- | --- |
| DataSet | Training | Validation |
| Adult DataSet | 50% | 50% |

For Predictive Modelling, we have implemented three below mentioned models:

**Logistic Regression:**

It is a statistical analysis model that helps to predict value on the basis of the prior observations of a data set. It suspects a dependent data variable that analyses the relationship between one or more existing independent variables. Hence, we have used Logistic Regression model to predict that the income of an individual exceeds than $50, 000. Below mentioned is the analysis of Logistic Regression:



Output file of Logistic Regression : <Logistic-Regression.lst>

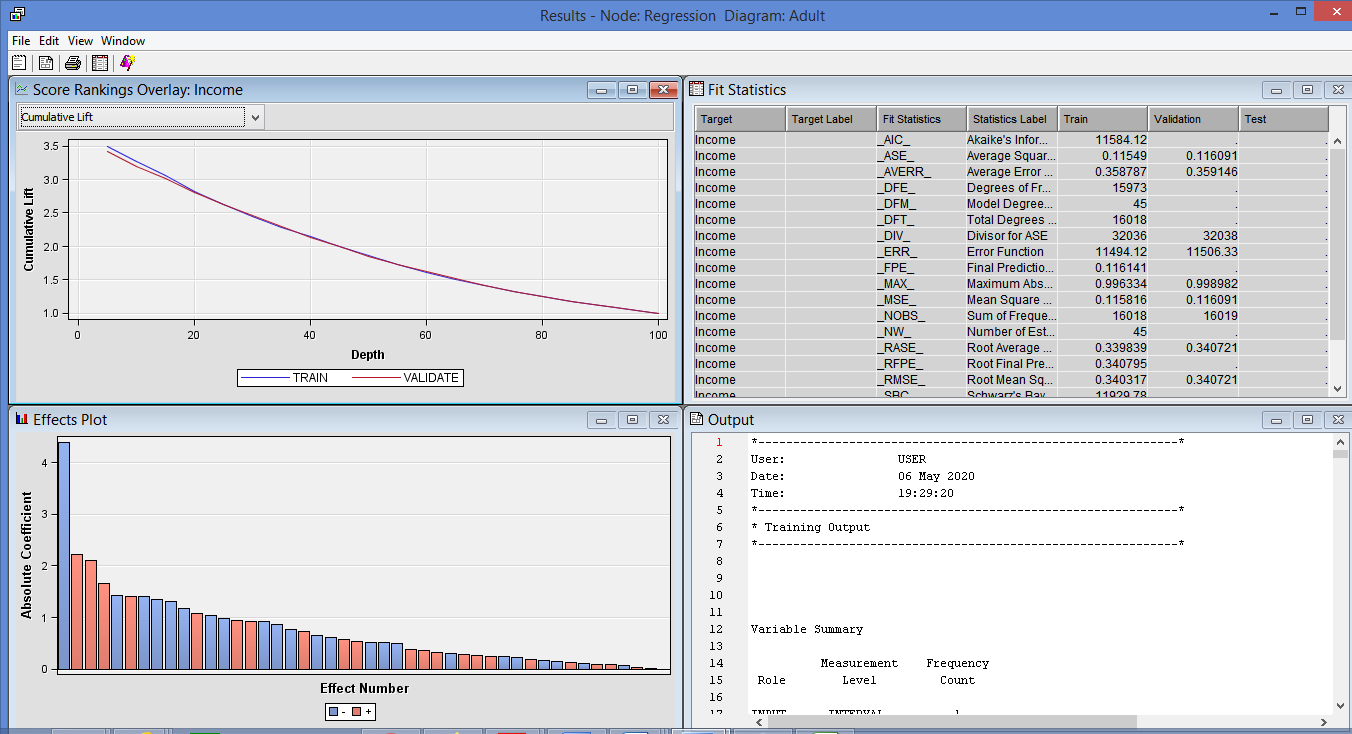
The Decision boundary is calculated as

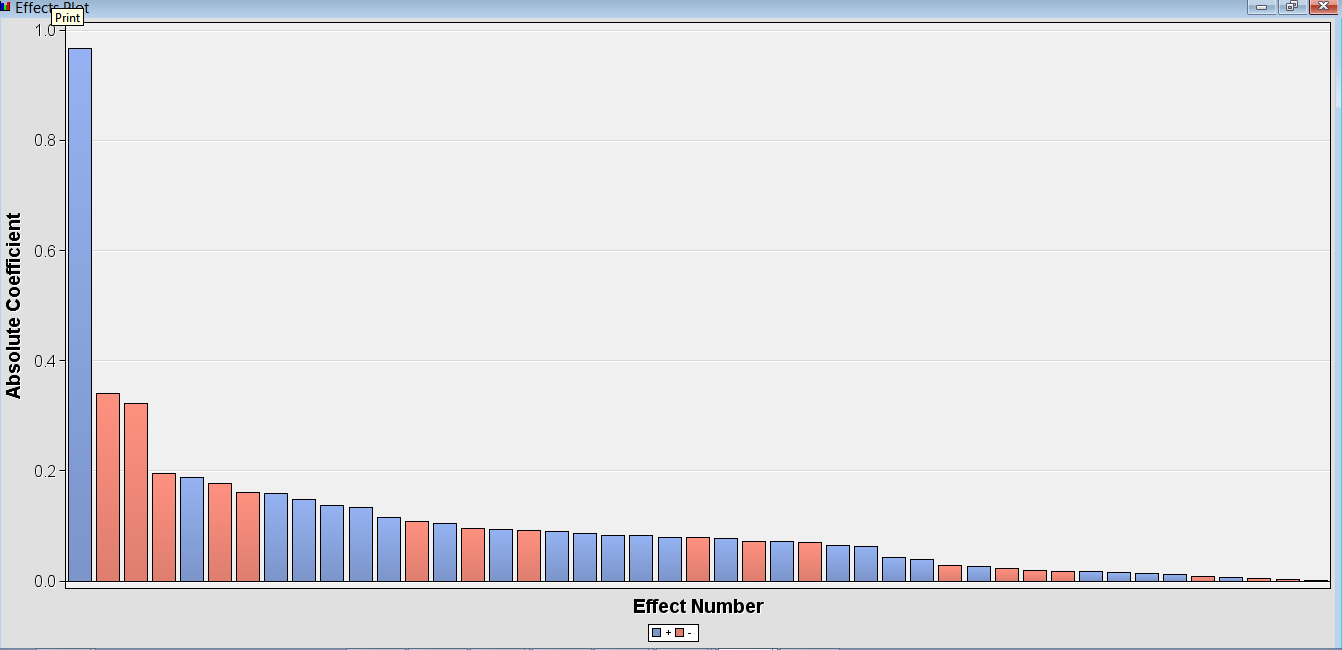
Income = Intercept+ EducationProf\_School+ educationDoctorate+ Marital\_stMarried\_Civ\_Spouse + Education5th and 6th --------------age+ relationshipHusband

However, it cannot be calculated for our graph, as there are 45 number of model degrees and there we specified some of the attributes in the above formula.

**Linear Regression:**

Linear model predicts two variables association and mathematical relationship by fitting a linear equation which is used to observe the data. The two variable associations are given by a straight line and only has one variable which is independent. Hence, we have used Linear Regression as a part of predictive modeling that predicts the income of an individual exceeds than $50, 000. Below mentioned is the analysis of Regression:





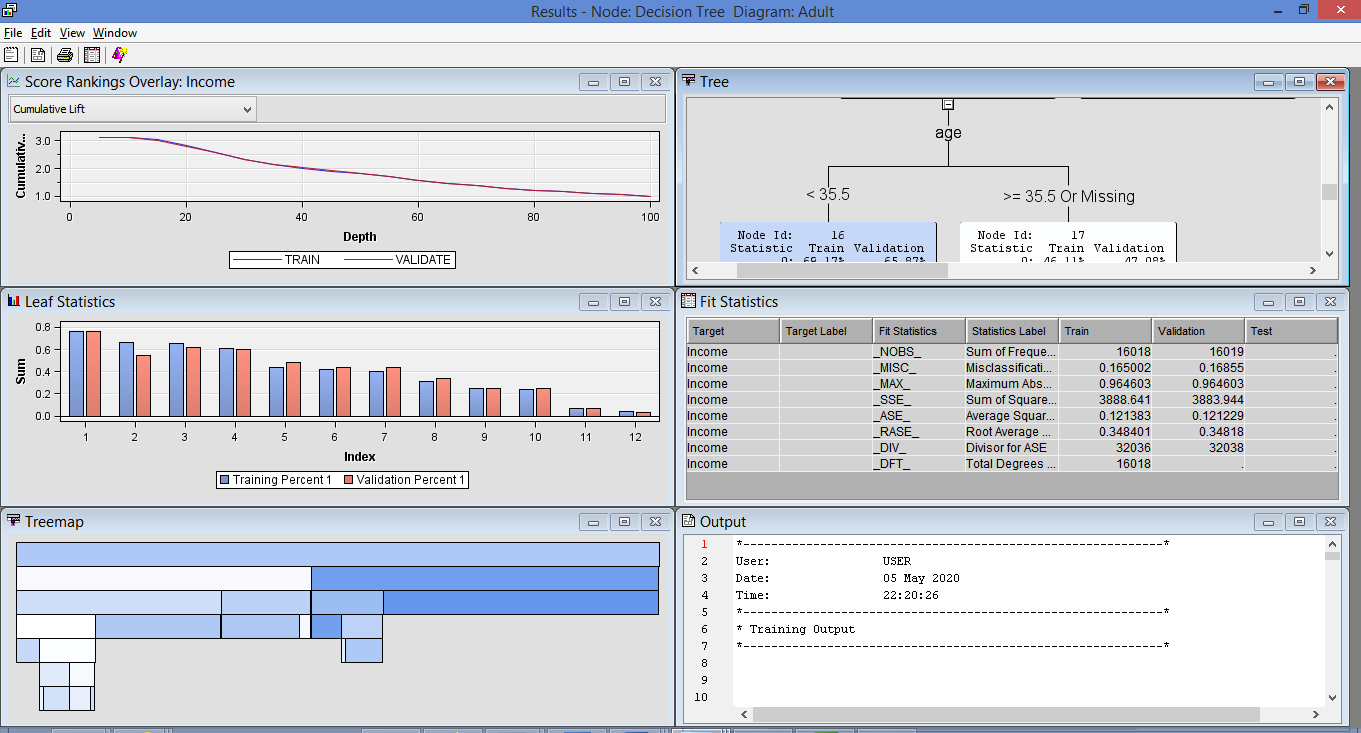
The above is effect plot and the equation of the linear regression model is given by :

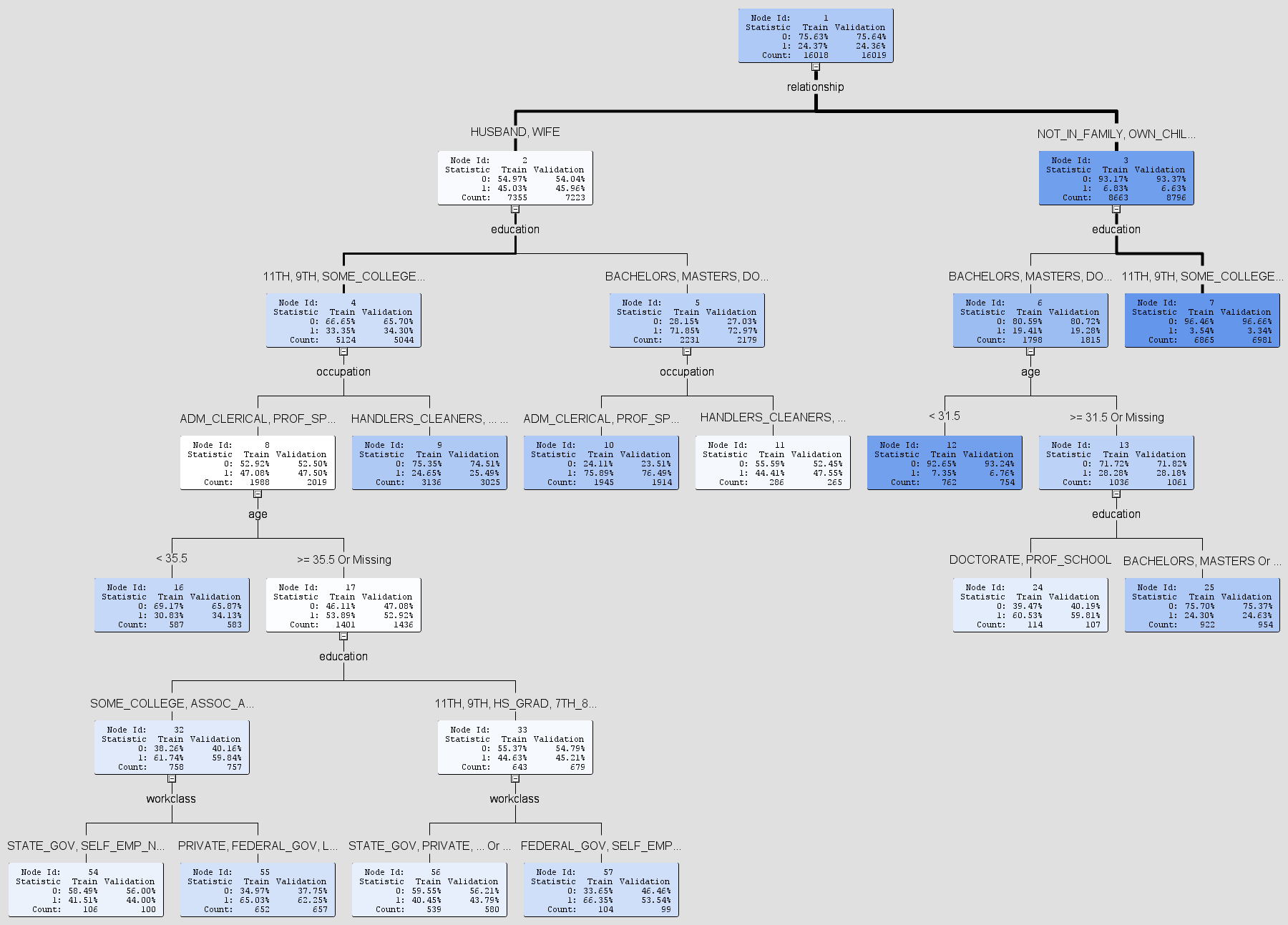
Income = Intercept+ EducationProf\_School+ educationDoctorate+ Marital\_stMarried\_Civ\_Spouse + Education5th and 6th --------------age+ relationshipHusband.

However, it cannot be calculated for our graph, as there are 45 number of model degrees and there we specified some of the attributes in the above formula

Output file of Linear Regression: [Outputs of SAS Code\Linear-Regression.lst](Outputs%20of%20SAS%20Code/Linear-Regression.lst)

**Decision Tree:** A Decision tree is a classification tree where the non-leaf node is termed as input feature. It aims to create a model which predicts the target attribute depending on several input variable. Therefore, we have used Decision Tree, to predict the actual target attribute for our data. The below is the analysis for Decision tree:





From the above decision tree it can be predicted that the most significant attribute is Relationship which is further splitting tree by education attribute and then followed by Occupation, Age and in the end followed by workclass.

The decision tree from SAS: [Outputs of SAS Code\Decision-Tree.bmp](Outputs%20of%20SAS%20Code/Decision-Tree.bmp)

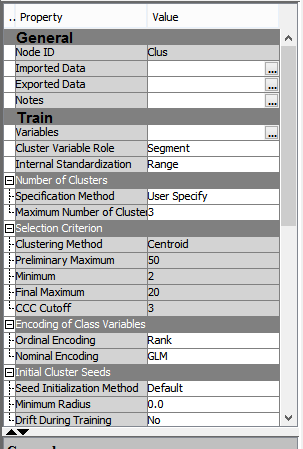
SAS Code for Decision Tree: [Outputs of SAS Code\SAS Code for Decision Tree.sas](Outputs%20of%20SAS%20Code/SAS%20Code%20for%20Decision%20Tree.sas)

* 1. **Descriptive Modelling:**

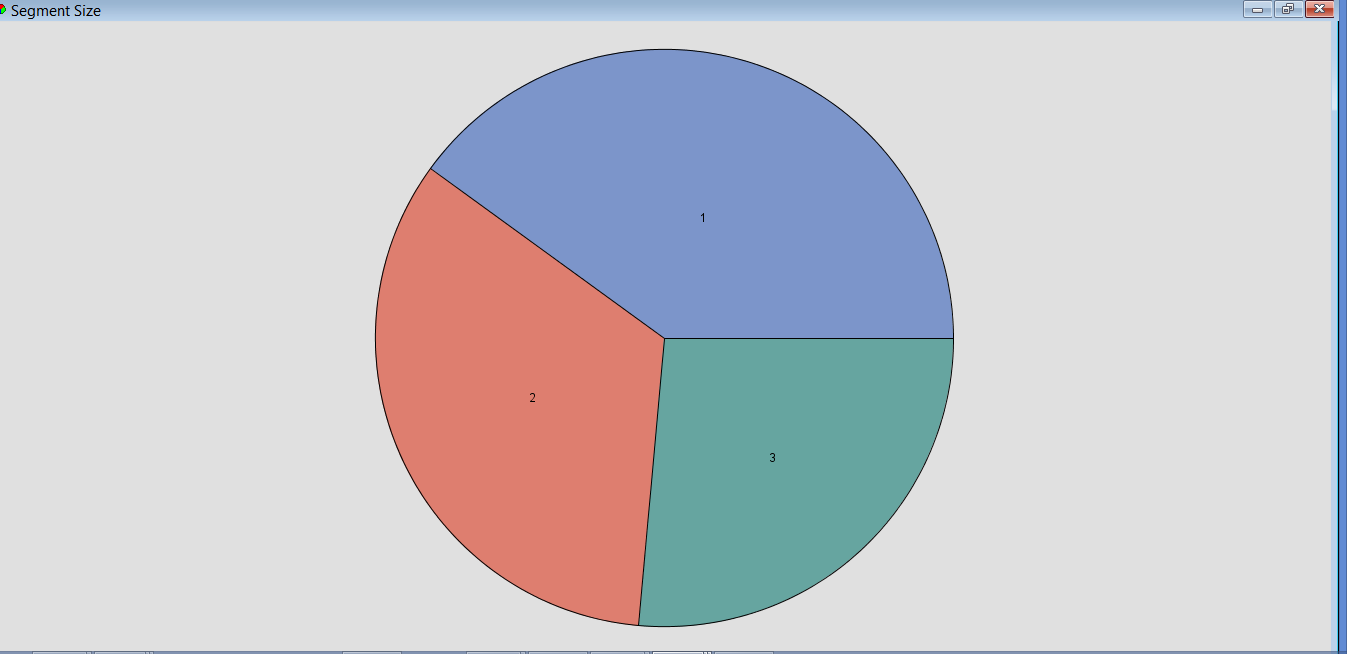
Descriptive modelling is used to detect the pattern and interpret the results by using various attributes that highlights the relationship in the data. The one most important guideline for this type of modelling is that there is no need of defining the target variable. Hence, for this I have used Cluster Analysis which includes K-means Clustering and Clustering. The analysis of both the model is given below:

1. **K-means Clustering (Range):**

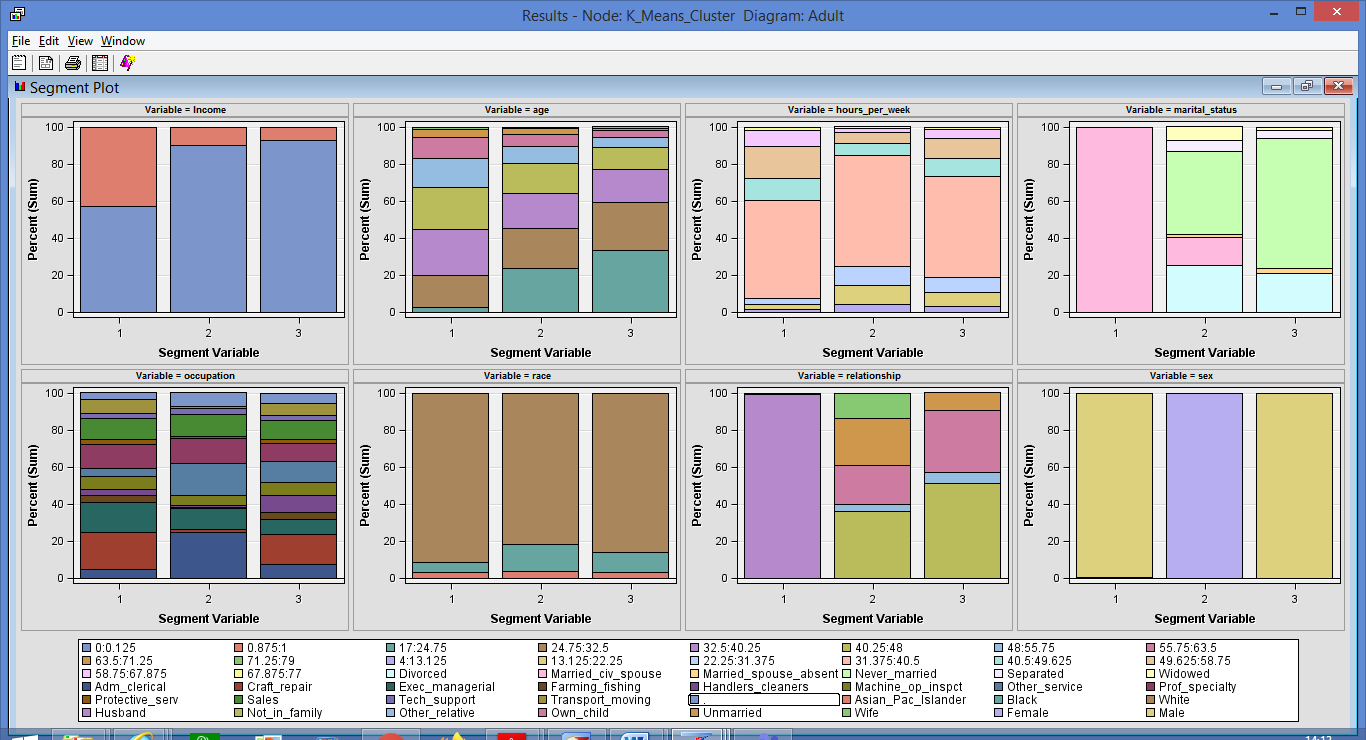
The K-means clustering algorithm is used to cluster the different observations in groups with related observations without any previous knowledge of those relationships. As a result, we have used K-means clustering to group the observations that are related to each other and the only differtiation is there should be no target variable in clustering. The analyses are as stated under**:**



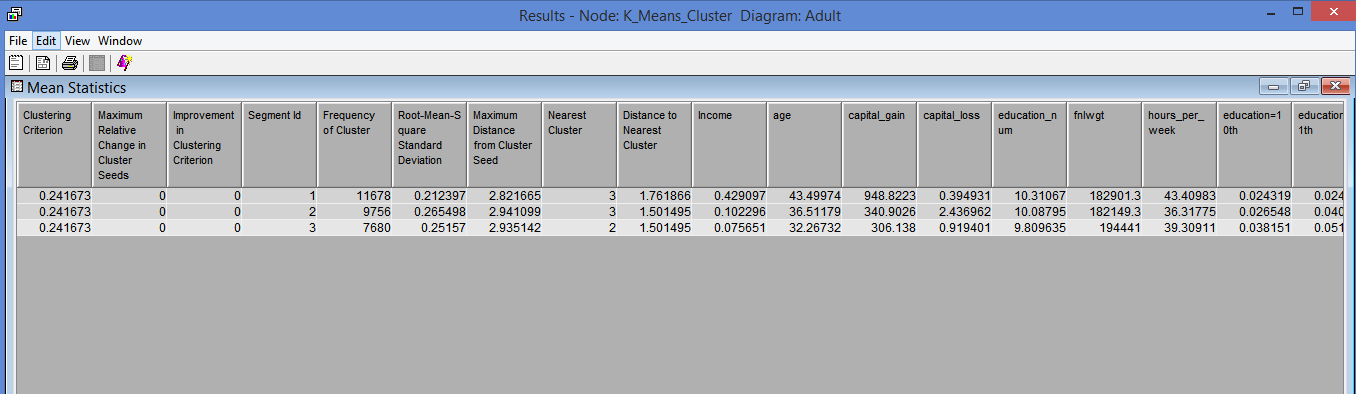
 For K-means clustering we have made the Internal Standardization as ‘Range’ and Specification Method as ‘User-Specify’ which means user can specify the number of clusters required for the analysis.

I

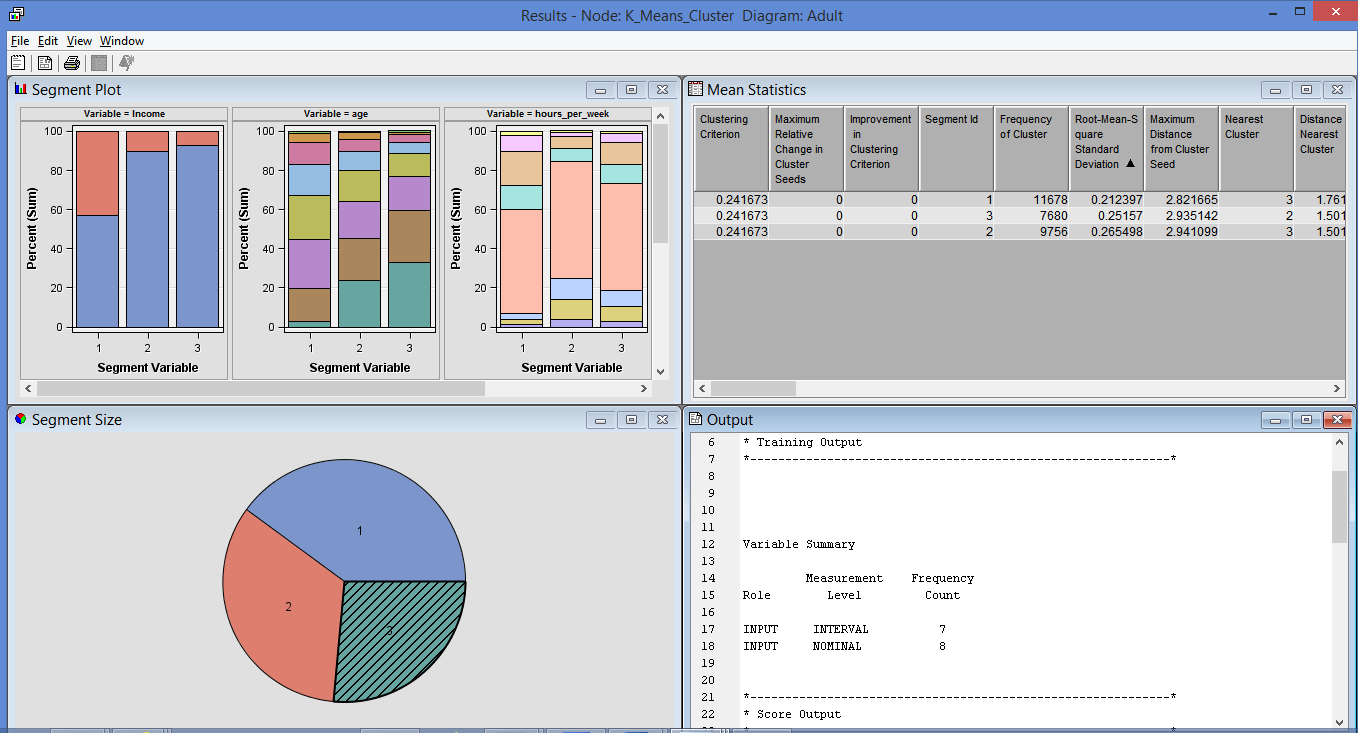
As the number of clusters specified is 3 and hence, there are three parts in the above pie chart.



The above is the segment plot for K-means clustering for which we have used some of the significant variables like age, income, education, occupation, relationship, sex workclass etc.



This is the mean statistics window in which we check for the values of Root-mean Square Standard Deviation.

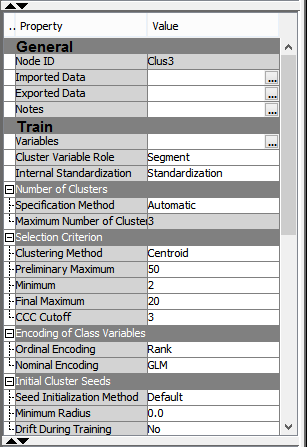


This is the entire K-Means clustering diagram for clusters.

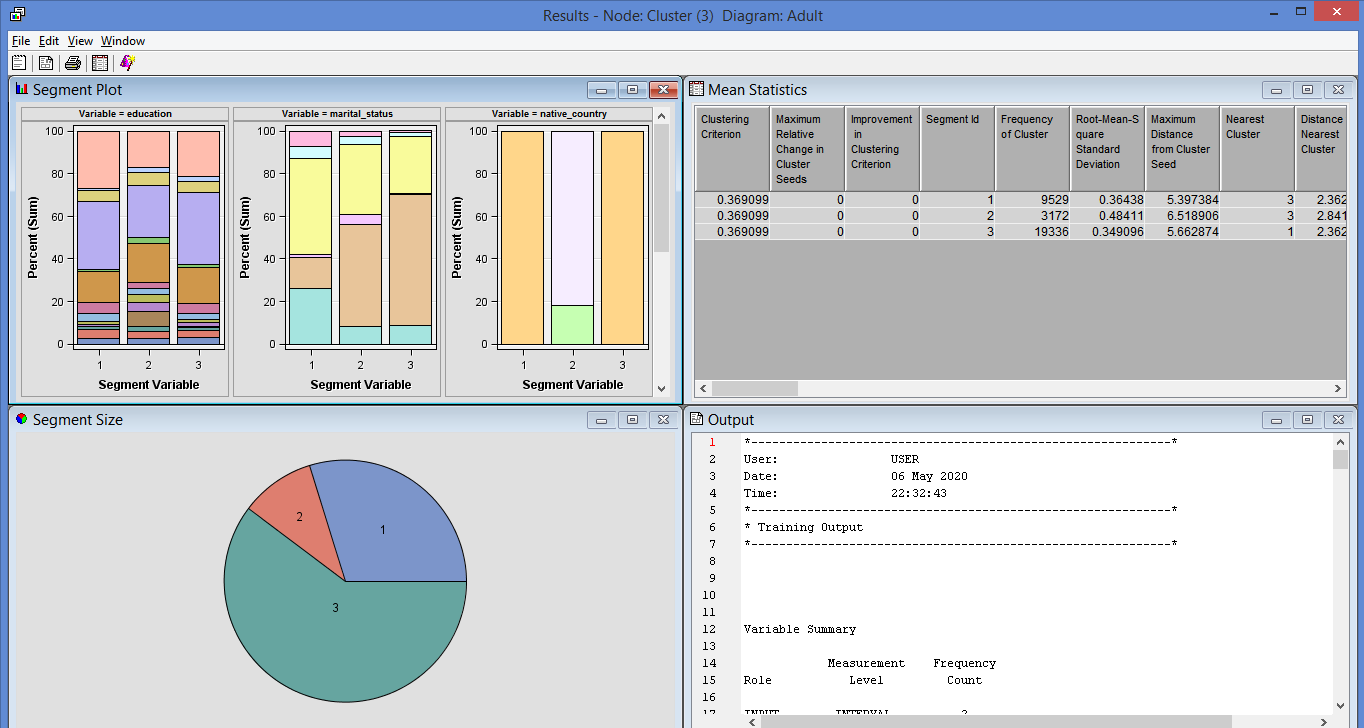
Node rules for K-means Clustering: [Outputs of SAS Code\Node Rules for K-means clustering.sas](Outputs%20of%20SAS%20Code/Node%20Rules%20for%20K-means%20clustering.sas)

1. **Clustering (Standardization):**

In this Clustering, we have used Internal Standardization as ‘Standardization’ which is used for normalisation to transform the input values by subtracting the mean and then dividing the standard deviation of the significant input varaibles. The analyses are as stated under:

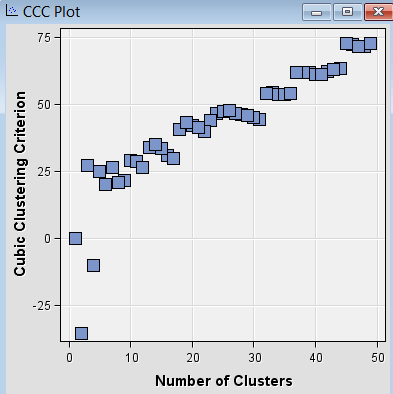


For clustering we have made the Internal Standardization as ‘Standardization’ and Specification Method as ‘Automatic’ which means user can specify the number of clusters required for the analysis.



The above is the segment plot for clustering for which we have used some of the significant variables like age, income, education, occupation, relationship, sex workclass etc.

**Analysis on the basis of Clusters:**



Node rules for clustering: [Outputs of SAS Code\Node Rules for clustering1.sas](Outputs%20of%20SAS%20Code/Node%20Rules%20for%20clustering1.sas)

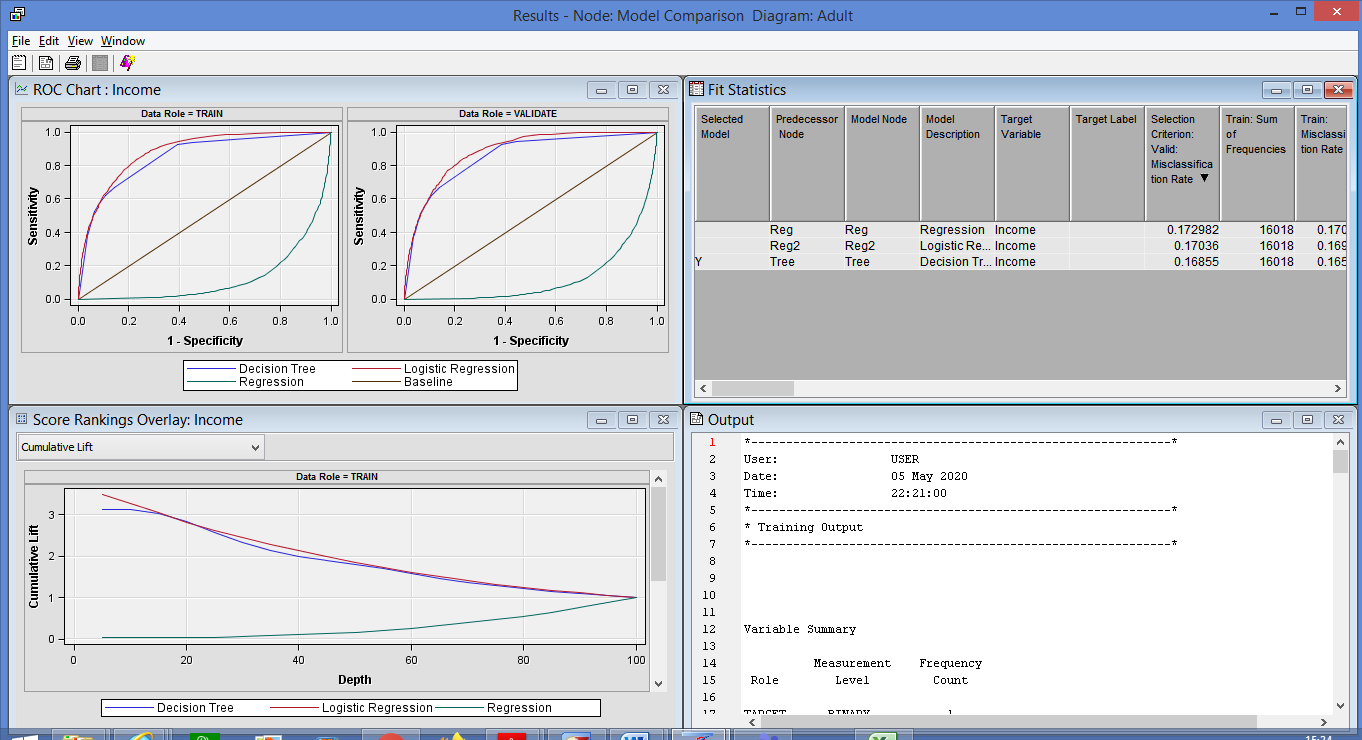
**7. Evaluation:**

While using CRISP model, in the phase of data understanding some attributes like Native-Country, Capital-gain, Capital-loss have single values for more than 80 % of the data and the Fnlwgt attribute was having the negative impact on our analysis, those were either normalizes or either removed which did not affected classifier performance.

Overall, the quality of data was good with very less occurrence of conflicting attributes and also, the data was cleaned from outliers and missing values in order to have the accuracy of the results.

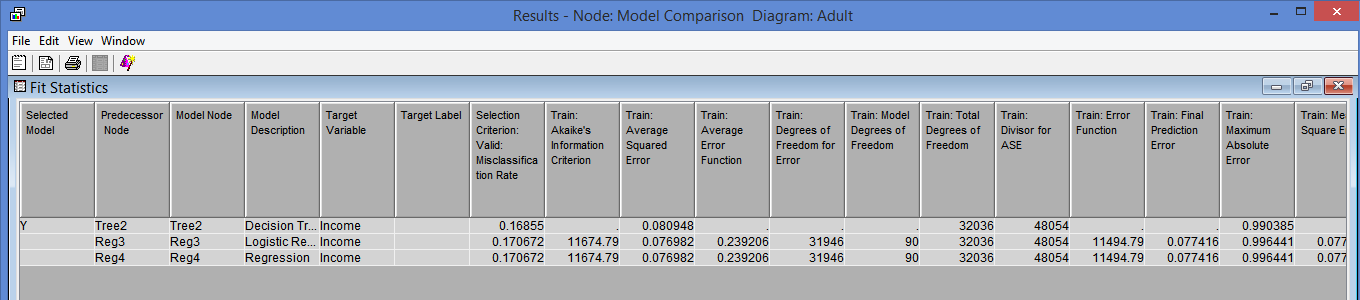
1. **1 Evaluation of Results for Predictive Modelling:**

For comparing the results for Predictive modelling in SAS we made use of Model comparison node and then than it for the results. The analysis and Results are as stated under:



This is score ranking overlay for Income all predictive models used, which states that it is almost same for both train and validation dataset.

Fit Statistics window: This is the output window which shows the misclassification rate for both Train and Validation data for all three predictive models.This fit statistics window states that Decision tree is the best predictive modelling for with the Misclassification rate as 0.165064 for train dataset, however, it is 0.16855 for validate dataset.



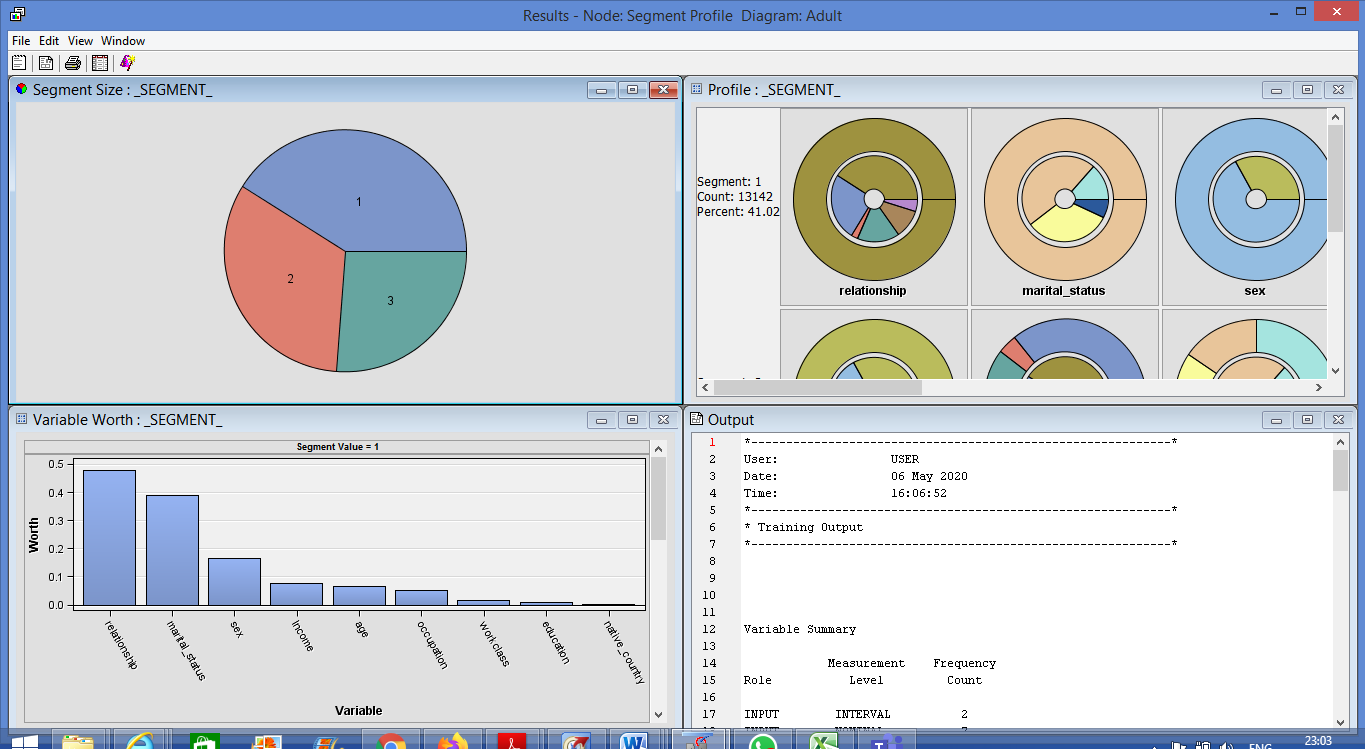
Output file Predictive modelling: [Outputs of SAS Code\Output file for Predictive modelling.lst](Outputs%20of%20SAS%20Code/Output%20file%20for%20Predictive%20modelling.lst)

* 1. **Evaluation for Descriptive modelling:**

For Descriptive modelling, we have used Segment Profile node along with Cluster node in SAS. The analysis and Results are stated under:

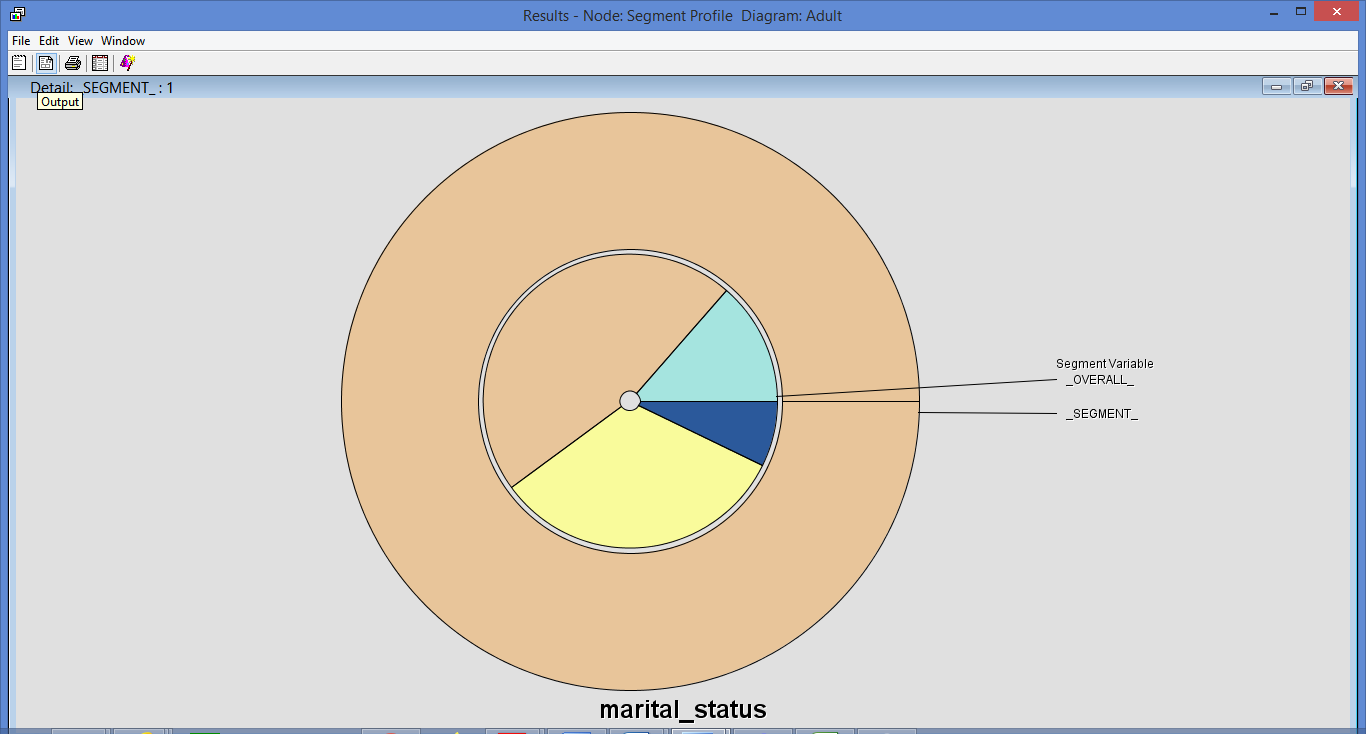
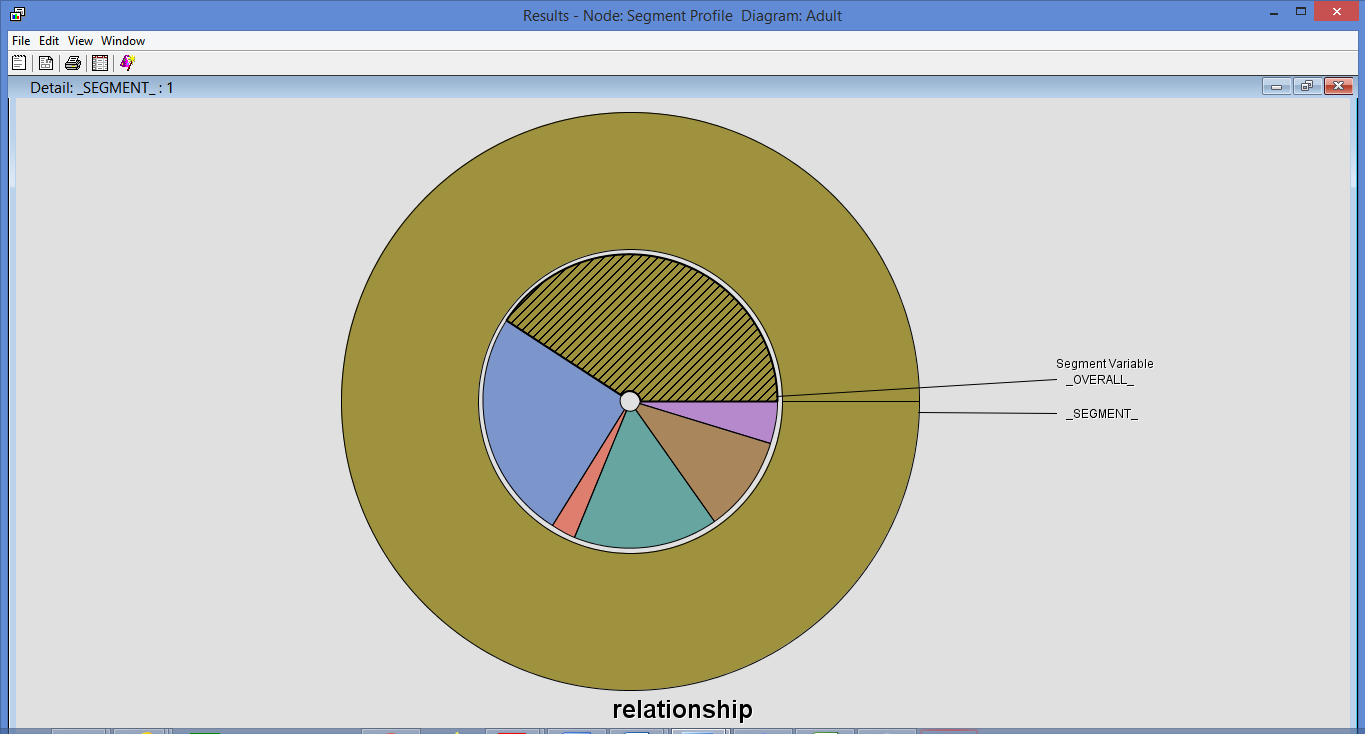
**K-Means Clustering:**

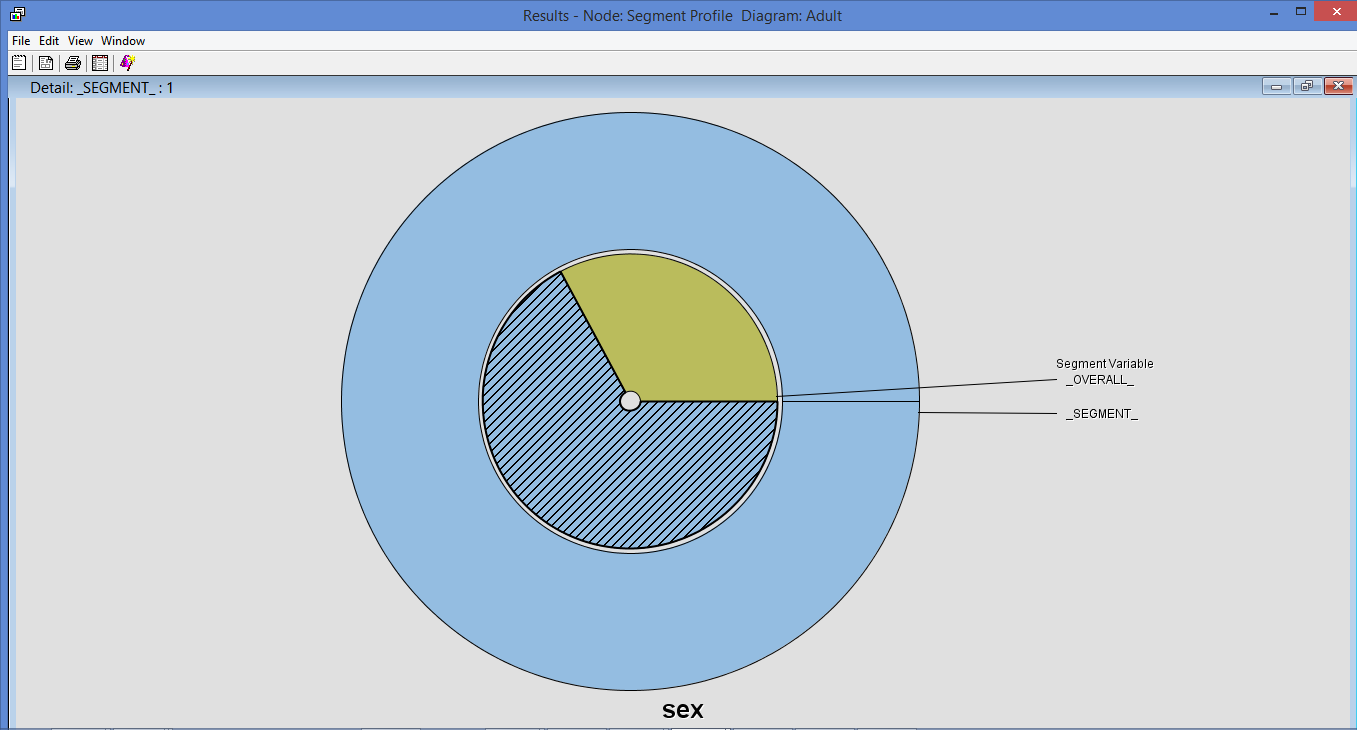
The segment size is 3



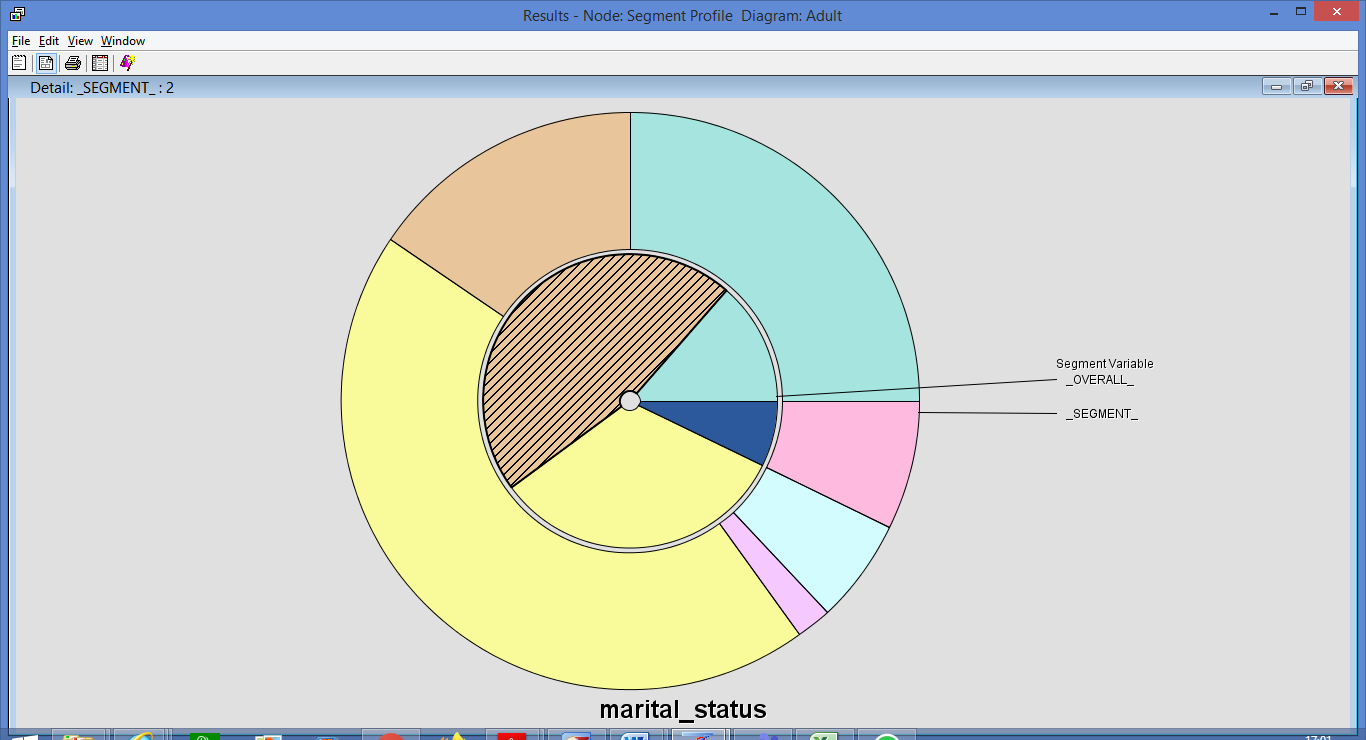
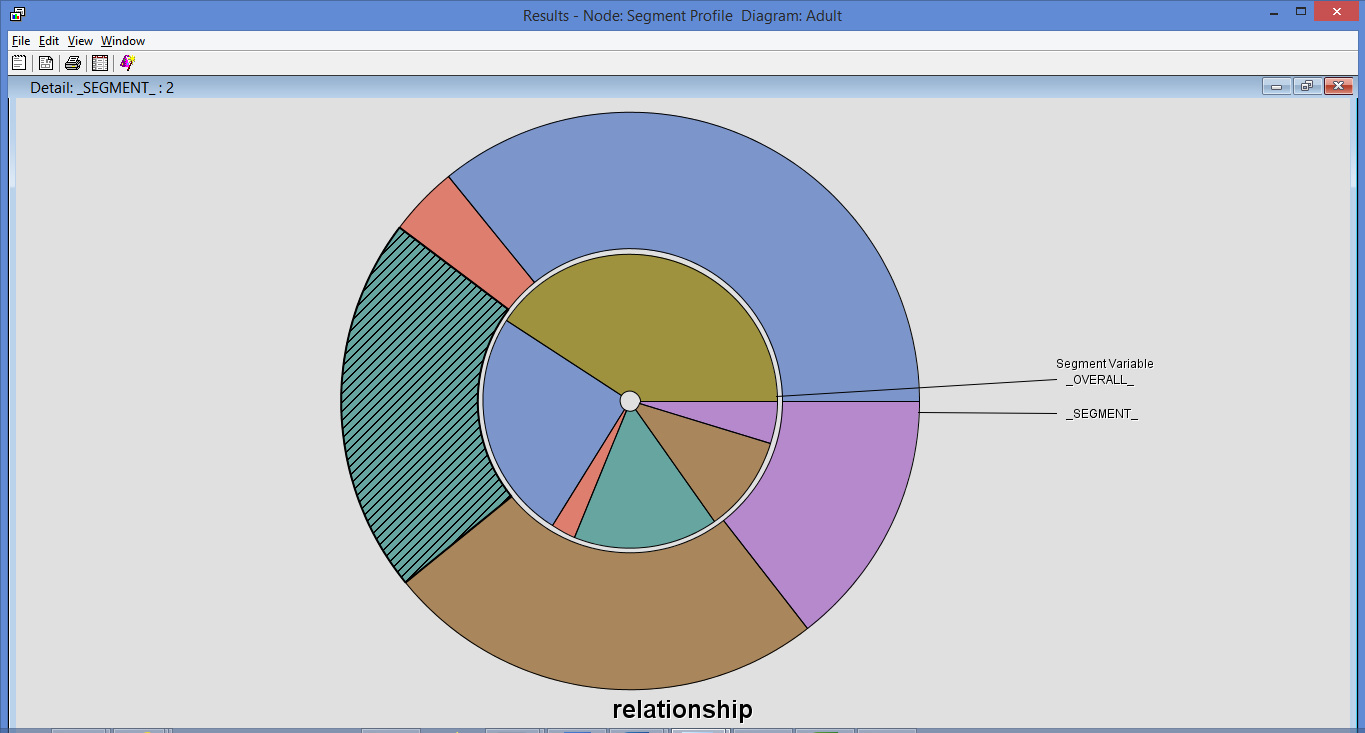
The variable worth window which states that Relationship and Marital status is the most significant variable for all three segments and Native-country is the least significant.

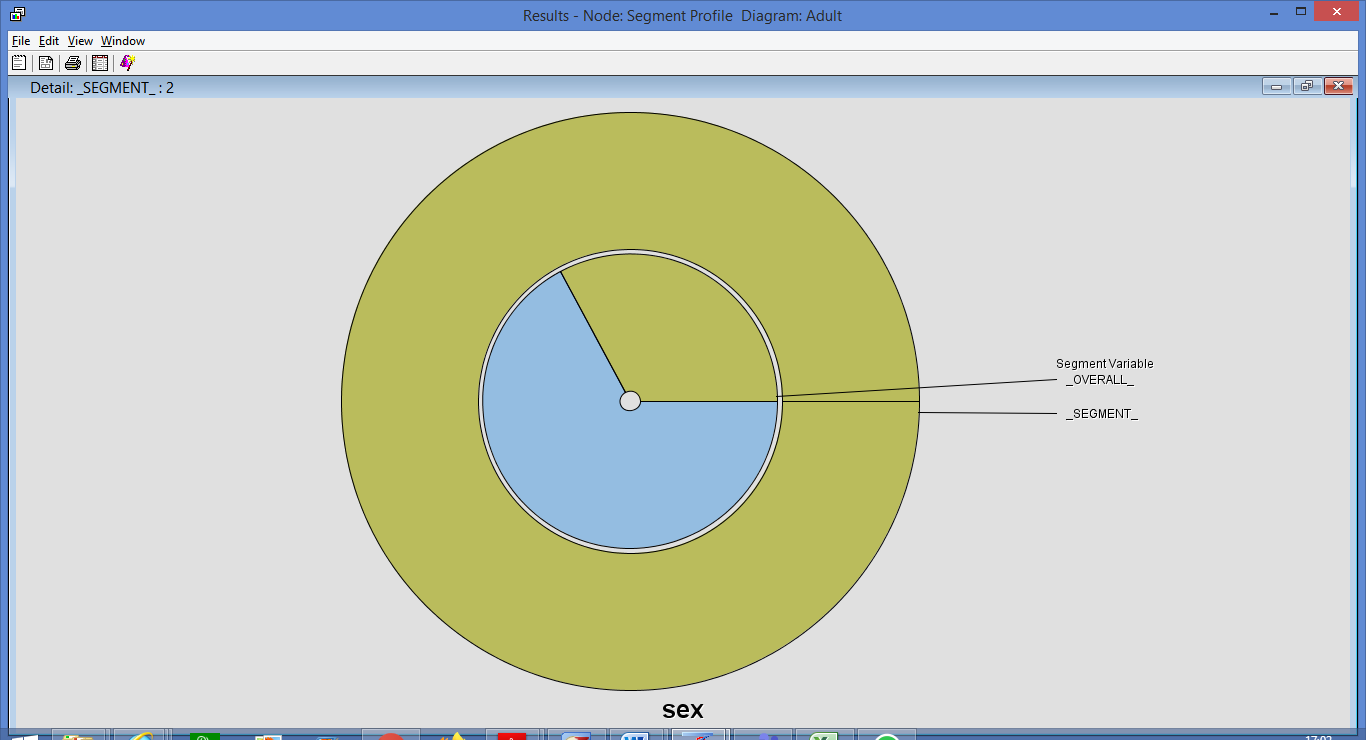
From Segment 1: Those who are husbands or not in a family are either married with civic spouse or are never-married are mostly males with 67.14736% have income greater than $50, 000





From Segment 2: Those who are not in a family or Unmarried (relationship status) are either never-married or divorced (marital status) are mostly males (sex) with 67.14736% who have income greater than $50, 000

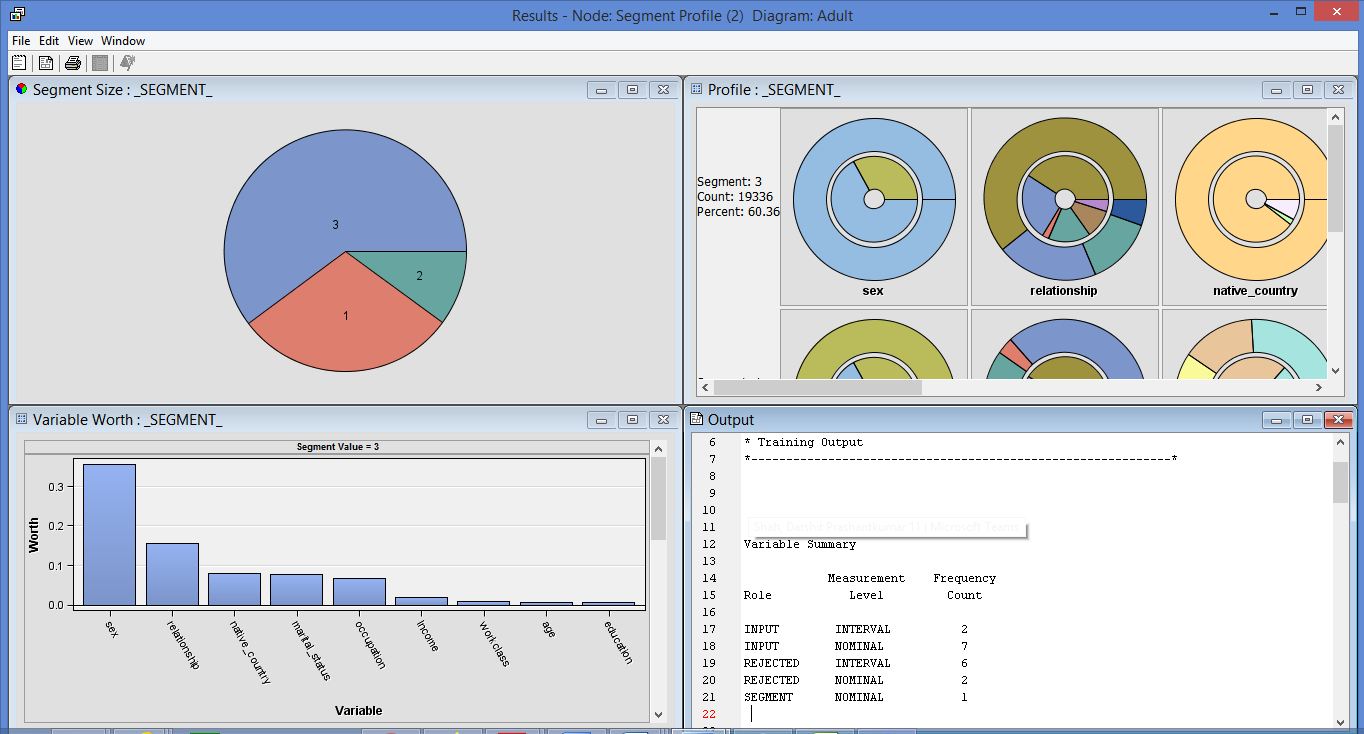




Output file of Segment Profile: [Outputs of SAS Code\Output file for K-Means Clustering segment modelling.lst](Outputs%20of%20SAS%20Code/Output%20file%20for%20K-Means%20Clustering%20segment%20modelling.lst)

**Clustering (Standardization):**

The segment size is 3 and hence, there are 3 clusters:

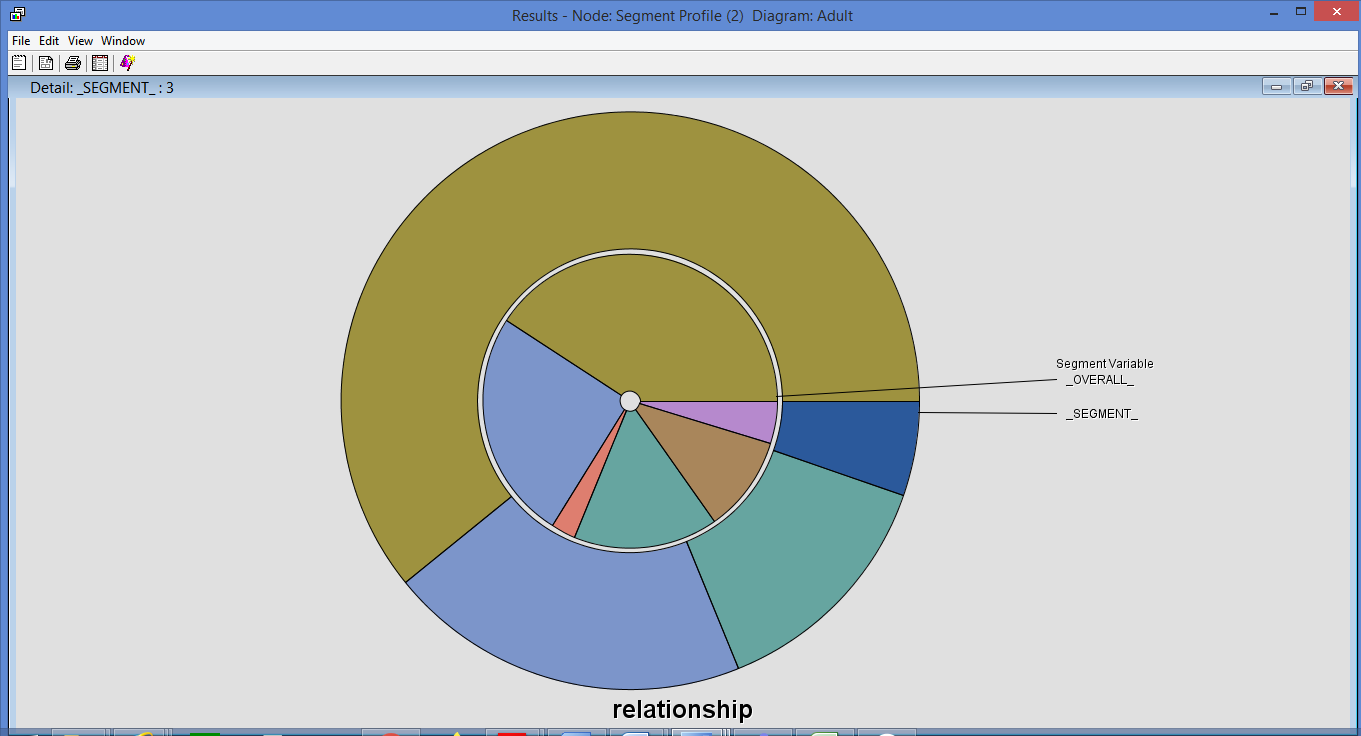
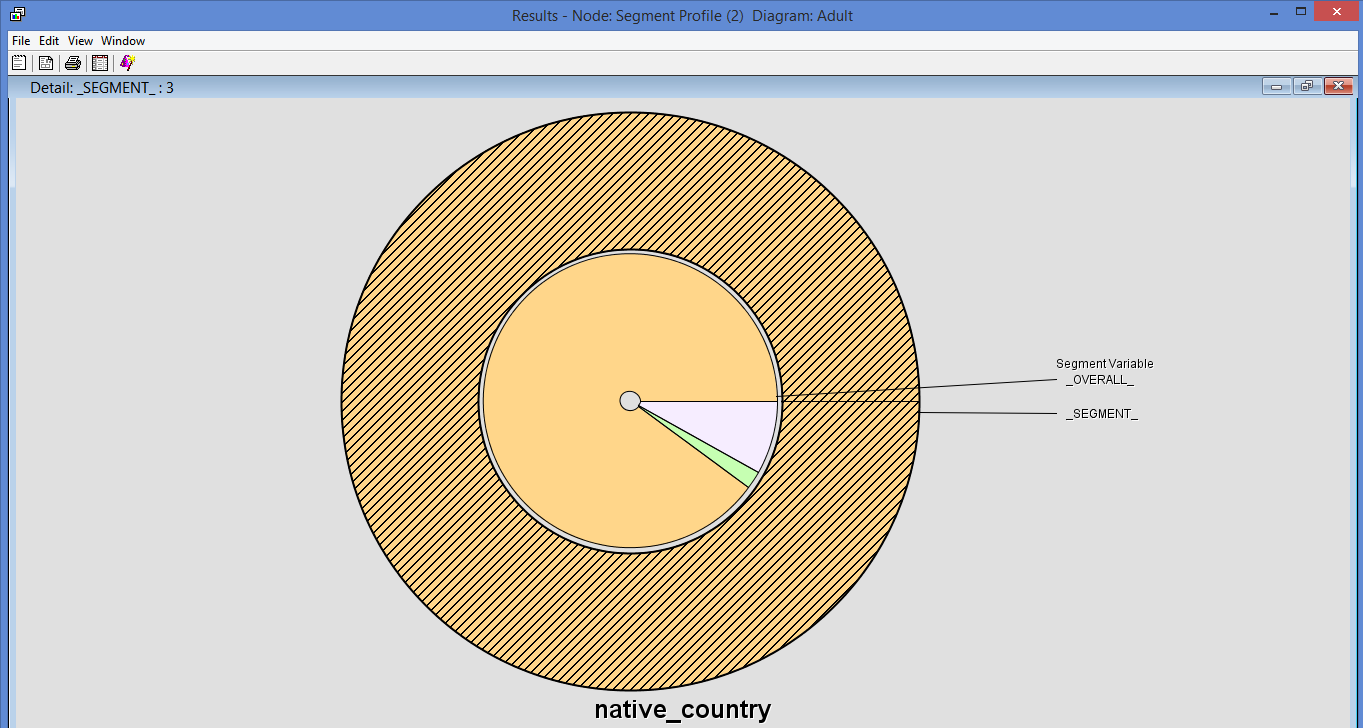


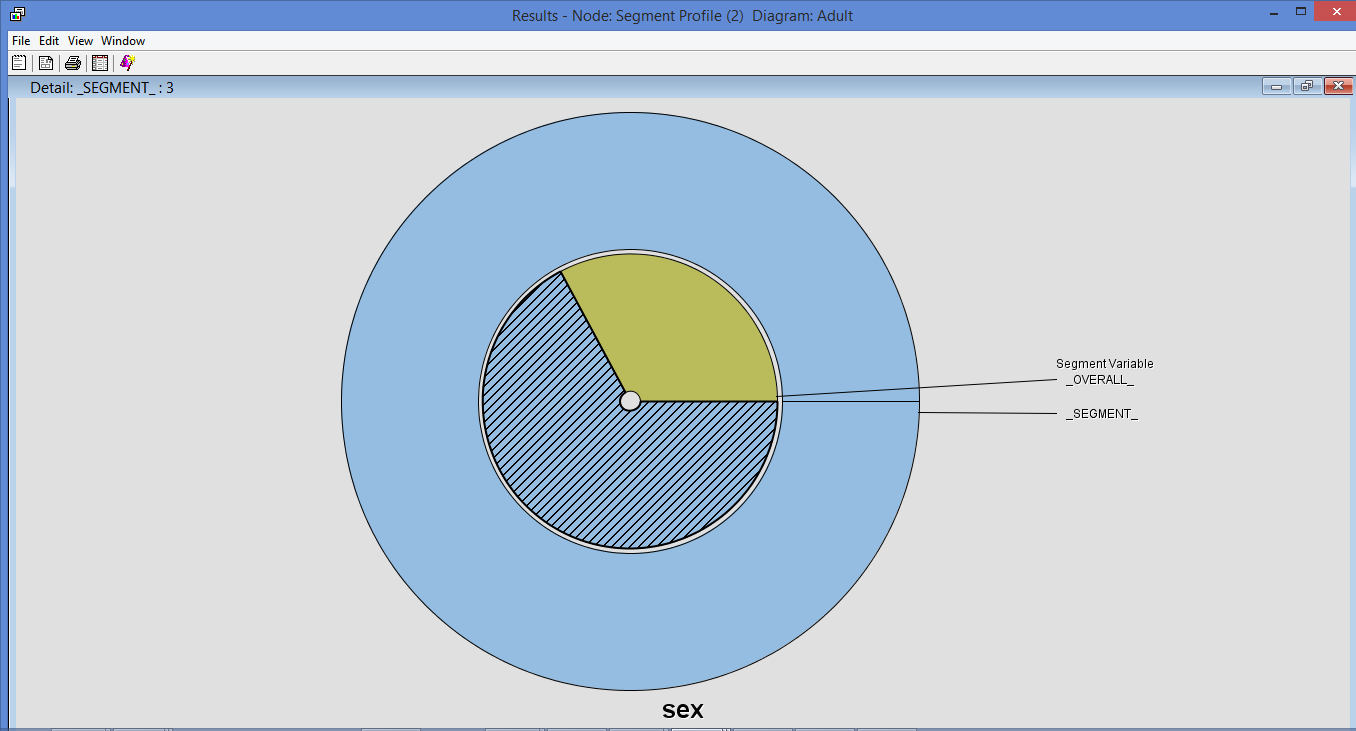
The variable worth window states that Sex and Native country is the important variable for all three segments.

Each segment interprets and which is as stated under:

**Segment 3:**

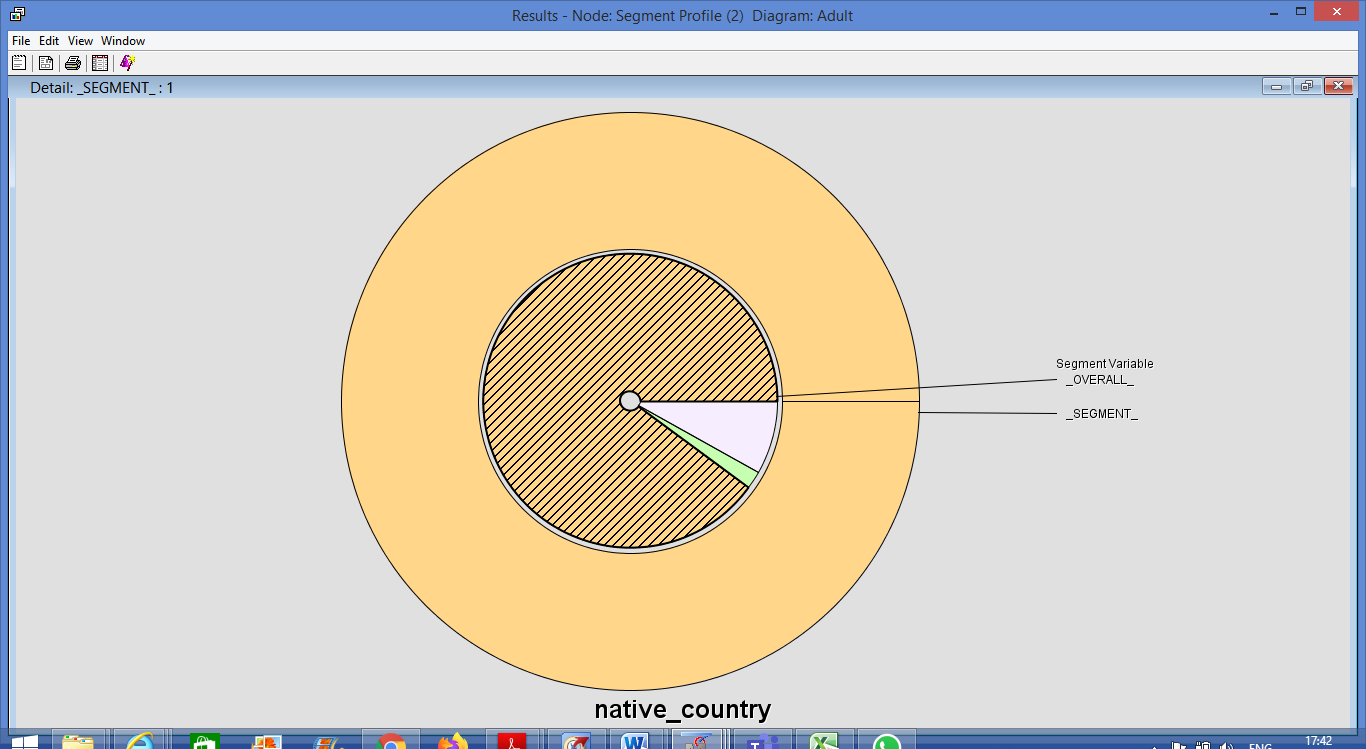
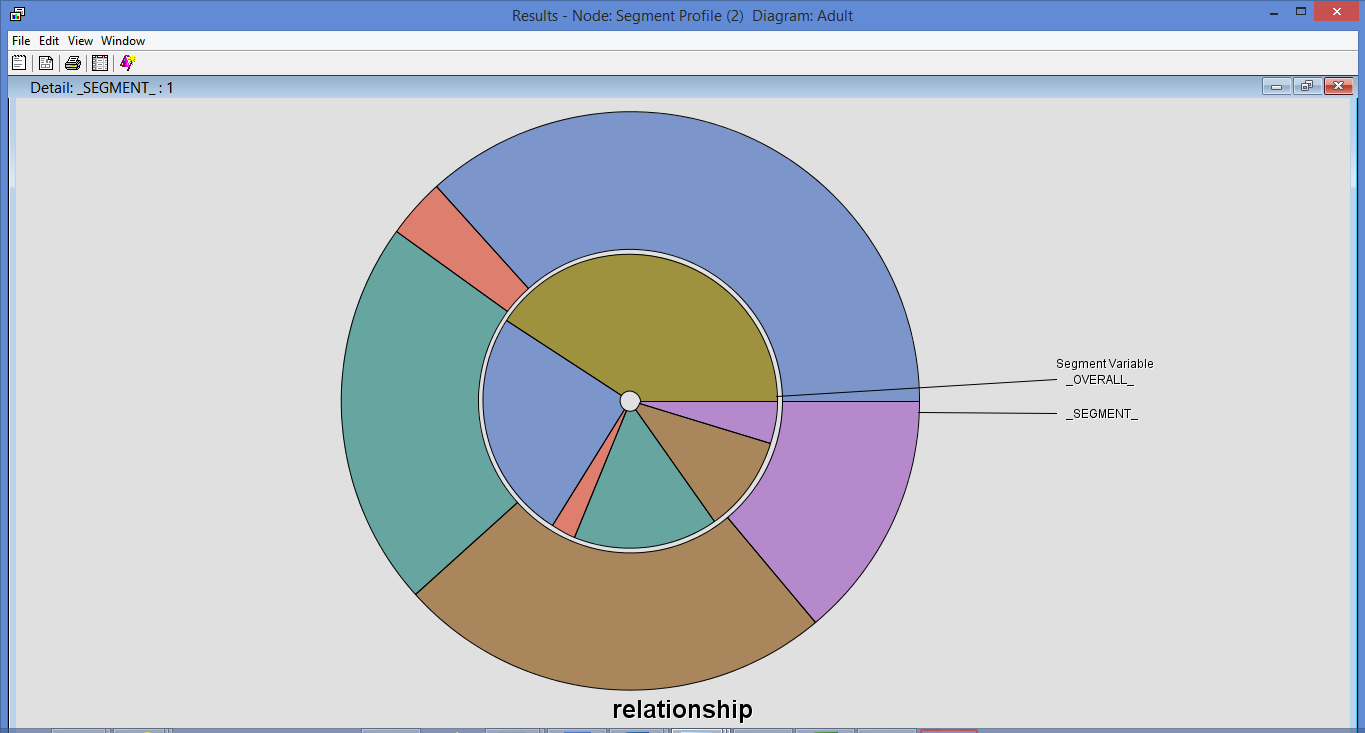
Around 90% of Individuals whose native-country is United States (native country) who are in relationship in either husband or not in a family (Relationship) are either married with civic spouse or never married (marital status) are males which are around 67.14736 earning more than $50, 000

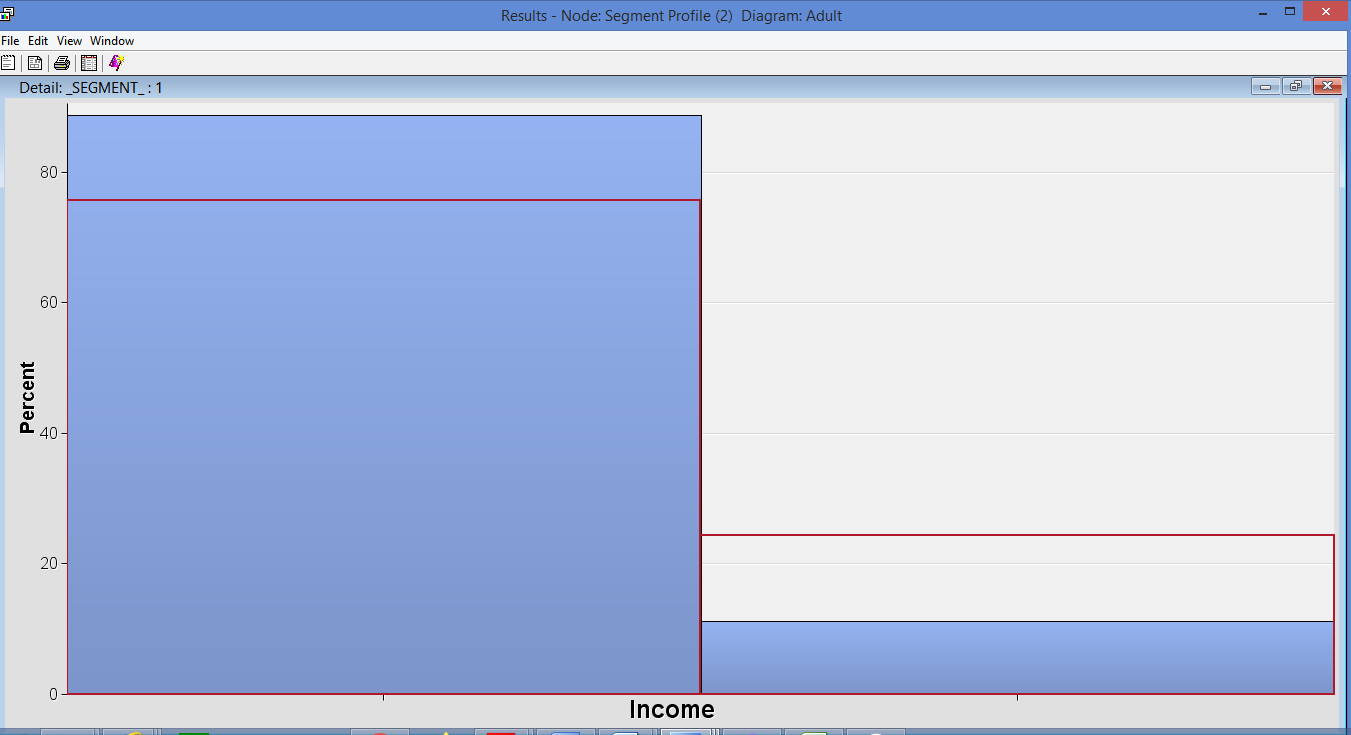




Segment 1:

Around 90% of Individuals whose native-country is United States(native country)who are in relationship in either not in a family or unmarried (relationship)are either never married or divorced are mostly males with 67.14736 earning more than $50, 000.



The output file for Segment profile for clustering: [Outputs of SAS Code\Output file forClustering segment modelling.lst](Outputs%20of%20SAS%20Code/Output%20file%20forClustering%20segment%20modelling.lst)

**8. Conclusion:**

This report proposed the use of Descriptive and Predictive modelling on Adult Census Data. Finally, the Validation Accuracy obtained is 86% , which by the best of our knowledge, has been the highest ever numeric accuracy achieved by any Income Prediction Model so far. There are interpretations done based on descriptive and predictive modelling which are as stated under:

1. From all the three predictive models used above, it can be concluded that the decision tree model is the best model for Adult Census Dataset with least misclassification rate as 0.165064
2. From the K-clustering It can be interpreted that:

i) Those whose income exceeds $50, 000 are either husbands or not in a family (relationship status) and are either married with civic spouse or are never-married (marital status) are mostly males with 67.14736%

ii) Those who are not in a family or Unmarried (relationship status) are either never-married or divorced (marital status) are mostly males (sex) with 67.14736% who have income greater than $50, 000

1. From Clustering it can be interpreted that:
2. Around 90% of Individuals whose native-country is United States (native country) who are in relationship in either husband or not in a family (Relationship) and are either married with civic spouse or never married (marital status) are males which are around 67.14736 earning more than $50, 000
3. Around 90% of Individuals whose native-country is United States(native country)who are in relationship in either not in a family or unmarried (relationship)are either never married or divorced are mostly males with 67.14736 earning more than $50, 000

Overall, it can be predicted that are around 67% males who have income of more than $50, 000 , however, females are around 23%

**9. References:**

[1] Kaya, F., 2008, 'Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers', University of Maryland publications.

[2] Kohavi, R., 1996, 'Scaling Up the Accuracy of Naïve Bayes Classifiers: a DecisionTree Hybrid', Proceedings of the second international conference on knowledge discovery and data mining.