Team Based Assignment

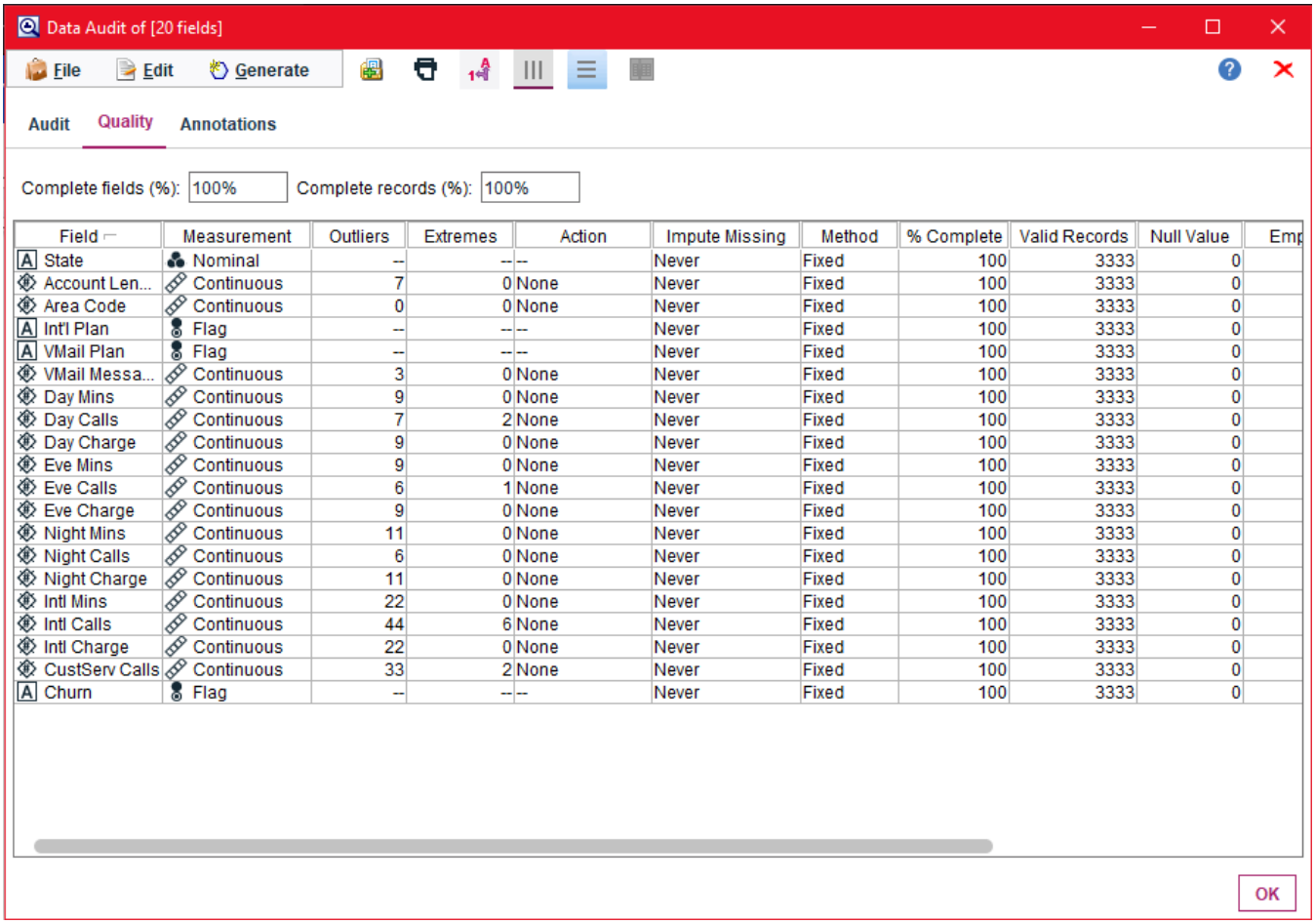
A) Use the churn data set for the following exercises.

1. Explore whether there are missing values for any of the variables.

After importing the source file 'Churn.xlsx' and reading the values in the data by name, connect with an output node as shown in the figure.

When the connected *data audit node* is executed or run, a data audit table is displayed. In this table the *Quality tab* gives the table as shown below.

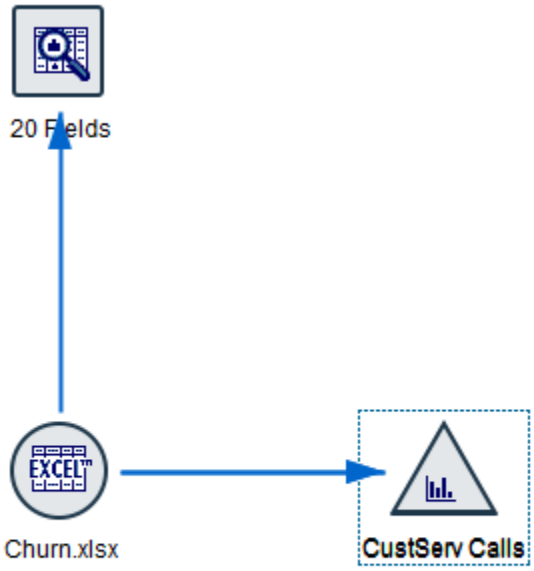
The *% Complete, Valid Records and Null Value* show that there are no missing values in the given file. Also, the valid values in the *Audit tab* show there are 3333 valid values, indicating zero missing values.

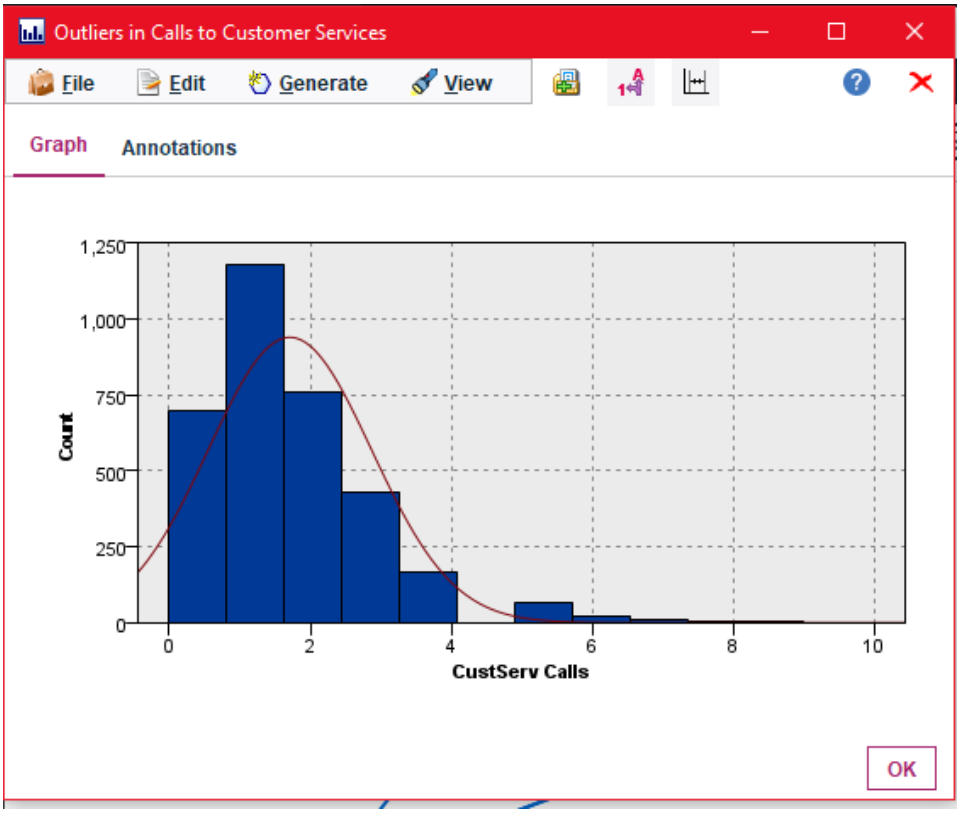


Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Emp
State	Nominal	--	--	--	Never	Fixed	100	3333	0	
Account Len...	Continuous	7	0	None	Never	Fixed	100	3333	0	
Area Code	Continuous	0	0	None	Never	Fixed	100	3333	0	
Int'l Plan	Flag	--	--	--	Never	Fixed	100	3333	0	
VMail Plan	Flag	--	--	--	Never	Fixed	100	3333	0	
VMail Messa...	Continuous	3	0	None	Never	Fixed	100	3333	0	
Day Mins	Continuous	9	0	None	Never	Fixed	100	3333	0	
Day Calls	Continuous	7	2	None	Never	Fixed	100	3333	0	
Day Charge	Continuous	9	0	None	Never	Fixed	100	3333	0	
Eve Mins	Continuous	9	0	None	Never	Fixed	100	3333	0	
Eve Calls	Continuous	6	1	None	Never	Fixed	100	3333	0	
Eve Charge	Continuous	9	0	None	Never	Fixed	100	3333	0	
Night Mins	Continuous	11	0	None	Never	Fixed	100	3333	0	
Night Calls	Continuous	6	0	None	Never	Fixed	100	3333	0	
Night Charge	Continuous	11	0	None	Never	Fixed	100	3333	0	
Intl Mins	Continuous	22	0	None	Never	Fixed	100	3333	0	
Intl Calls	Continuous	44	6	None	Never	Fixed	100	3333	0	
Intl Charge	Continuous	22	0	None	Never	Fixed	100	3333	0	
CustServ Calls	Continuous	33	2	None	Never	Fixed	100	3333	0	
Churn	Flag	--	--	--	Never	Fixed	100	3333	0	

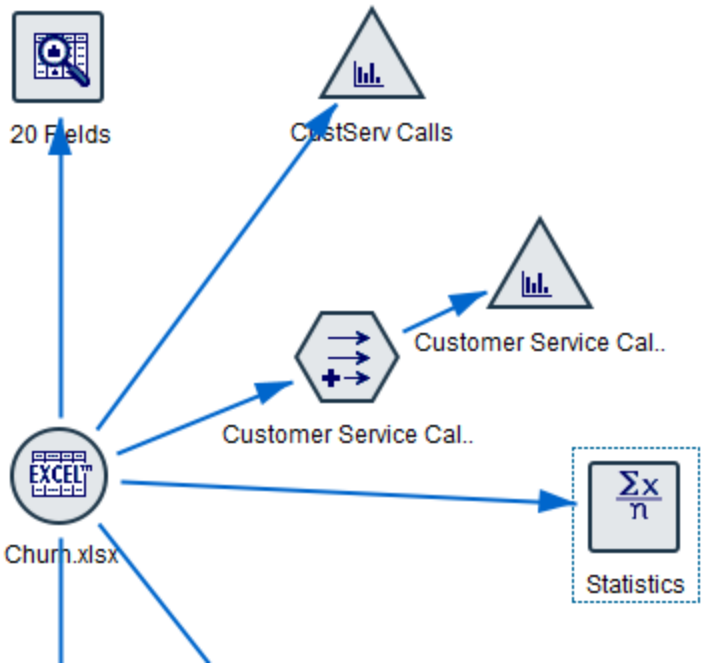
2. Use a graph to determine visually whether there are any outliers among the number of calls to customer service.

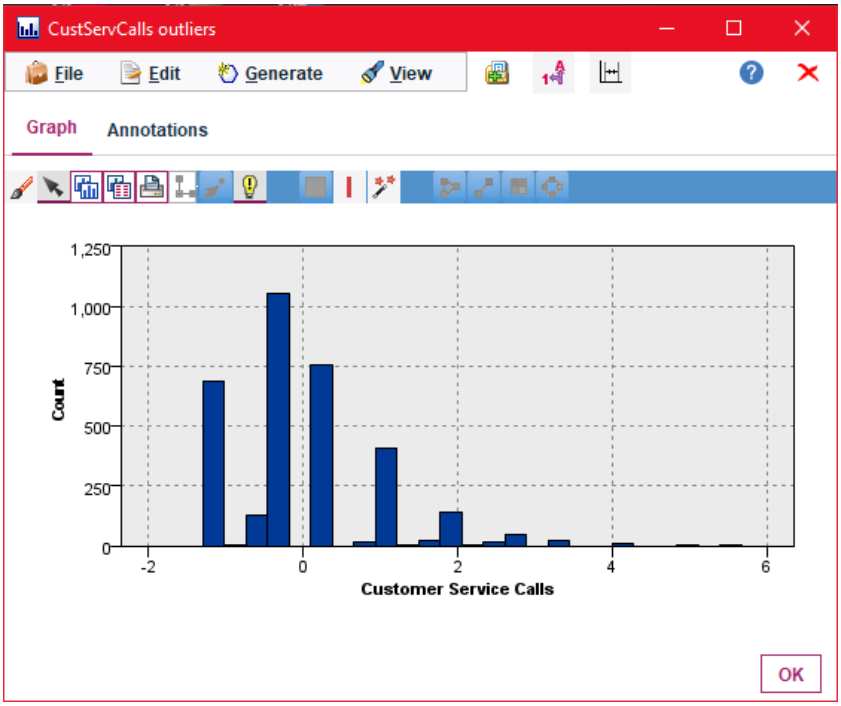
When a *histogram* is added as an output of the source file and executed as below, a graph of histogram is displayed. To clearly identify the outliers, a normalized curve is added, and bins are proportionately decreased.





Though the above graph visually shows the outliers, mathematically, the following execution of generating a *statistic node* and deriving a z score for the *customer service calls* attribute gives the histogram output.

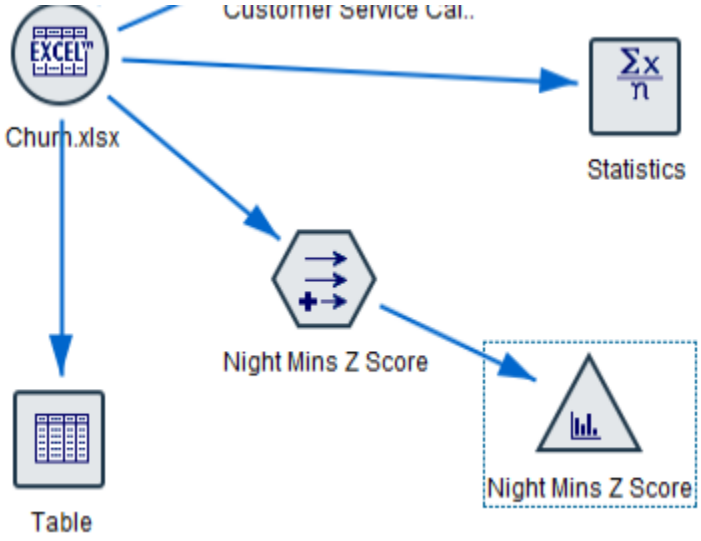

20 Fields
CustServ Calls
Customer Service Cal..
Customer Service Cal..
Churn.xlsx
Statistics


CustServCalls outliers
File
Edit
Generate
View
Graph
Annotations
1,250
1,000
750
500
250
0
Count
-2
0
2
4
6
Customer Service Calls
OK

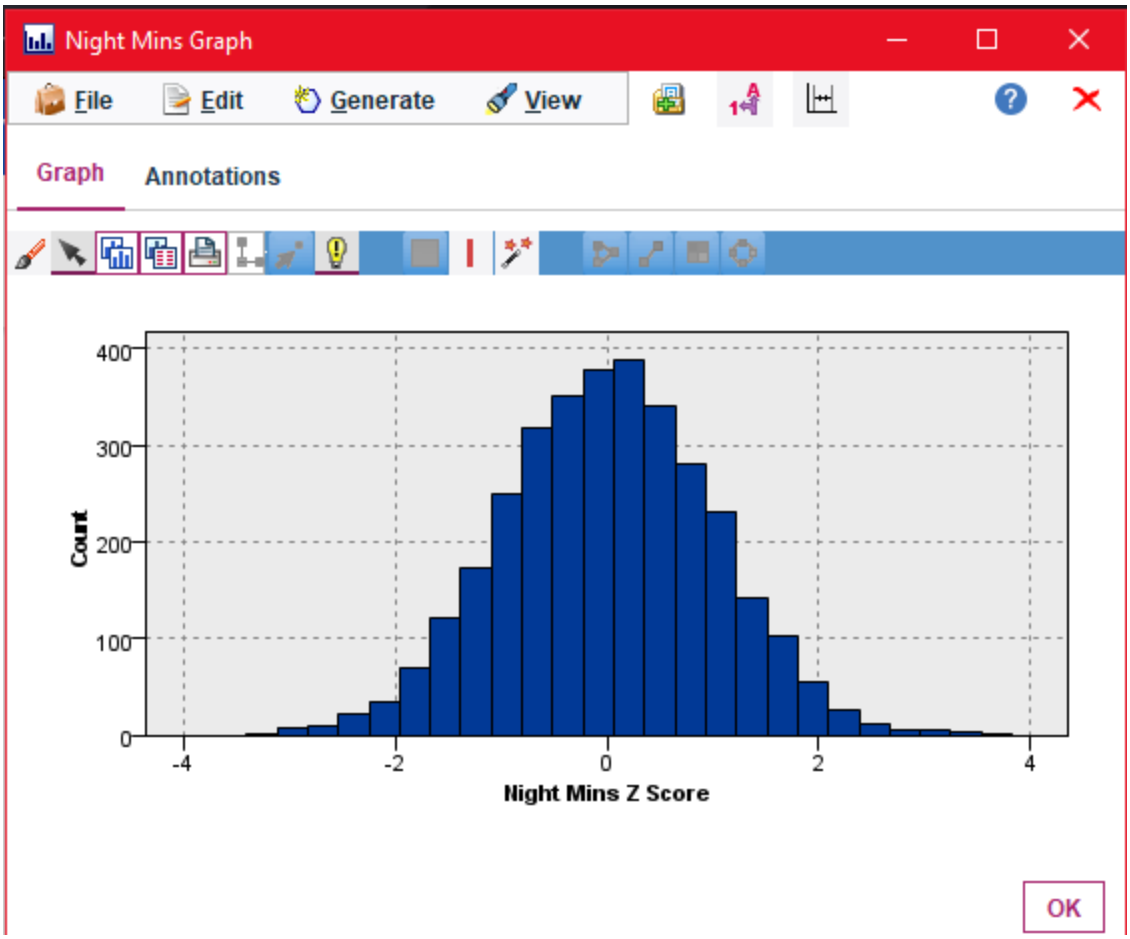
3. Transform the night minutes attribute using Z-score standardization. Using a graph, describe the range of the standardized values.

Hint: for 3, consider adding a Derive node to create each of the formulas (min-max normalization, and Z-score). You can find the required summary statistics (mean and std deviation) to calculate the Z-score using the statistics node

The z score to the *Night Mins* attribute is derived from the *derived node* after getting the results of mean and standard deviation from the *statistics node*.



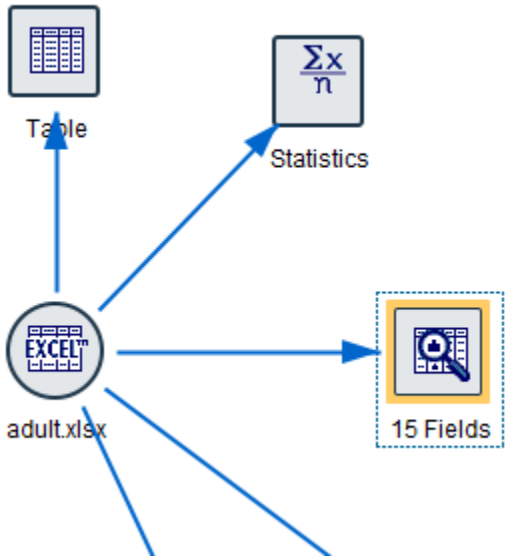
The derive node is then connected to a *histogram* to graphically give the range of standardized values of *night minutes*.


Night Mins Graph
File
Edit
Generate
View
Graph
Annotations
400
300
200
100
0
Count
-4
-2
0
2
4
Night Mins Z Score
OK

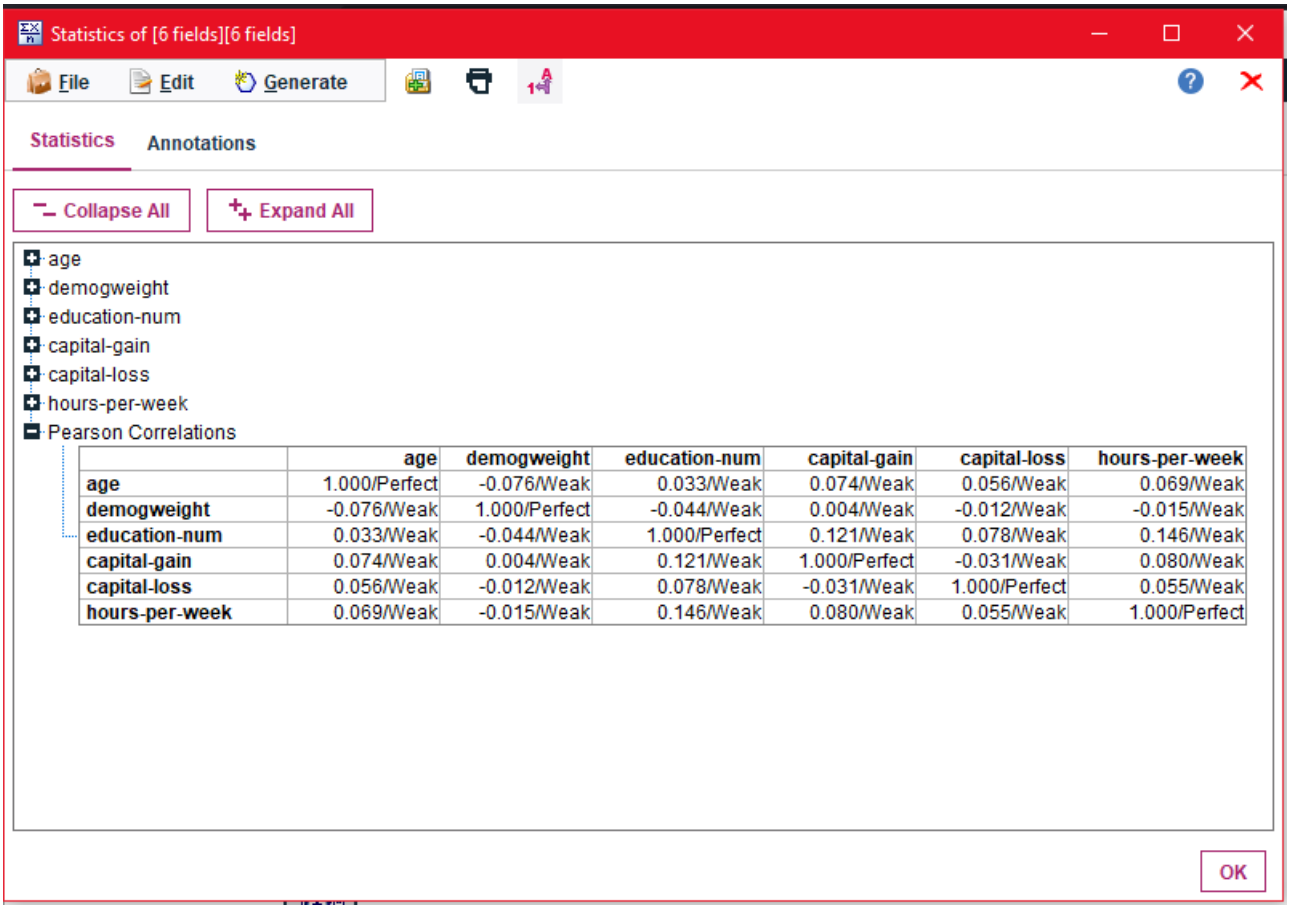
B) Use the adult data set for the following exercises. The target variable is income, and the goal is to classify income based on the other variables.

1. Investigate whether there are any correlated variables.

By connecting the *statistics node* to the source file 'adult.xlsx' a correlation is developed in this node with the other attributes in file as shown below.



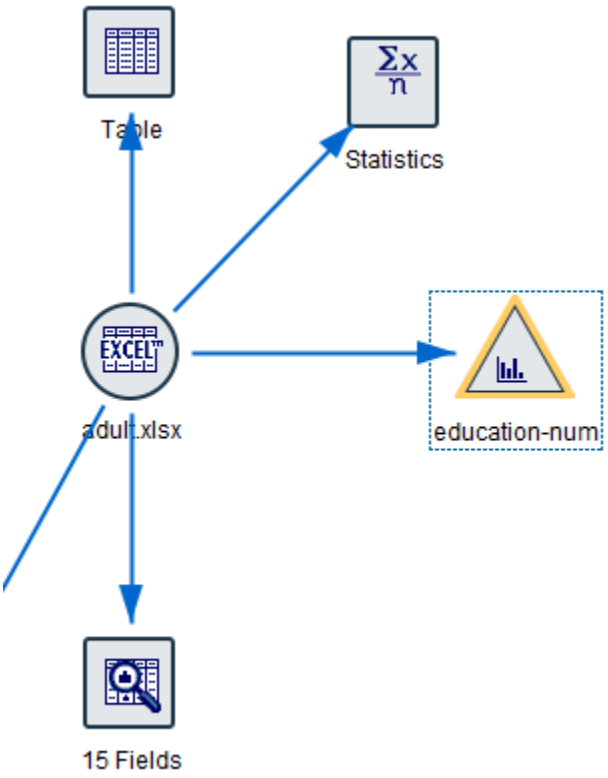
The following table of statistics is displayed where correlation of one attribute with the other is shown. It is clearly depicted that there is no correlation found between any attributes except of its own, which is an absolute correlation. Rather, a weak correlation is clearly found as the attributes, by definition, do not correlate with others.



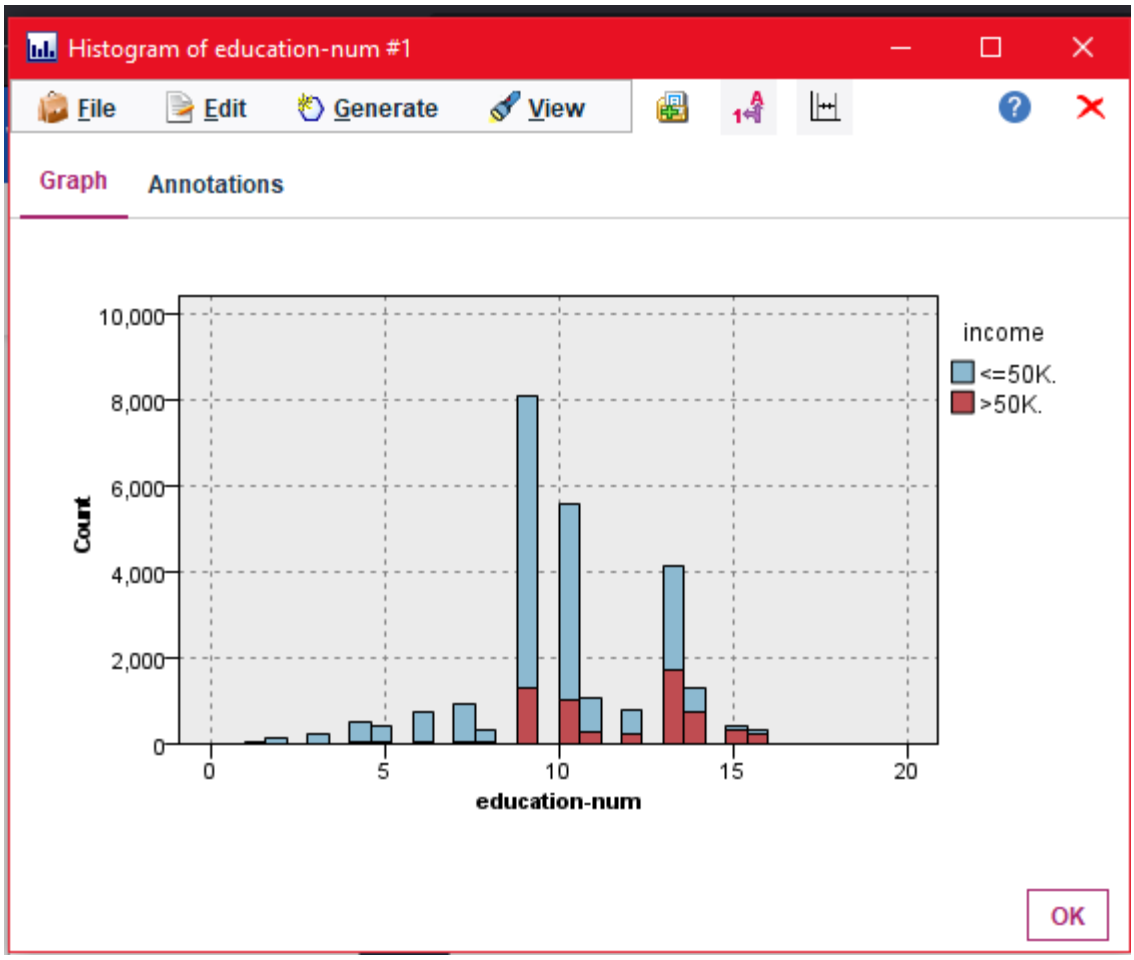
	age	demogweight	education-num	capital-gain	capital-loss	hours-per-week
age	1.000/Perfect	-0.076/Weak	0.033/Weak	0.074/Weak	0.056/Weak	0.069/Weak
demogweight	-0.076/Weak	1.000/Perfect	-0.044/Weak	0.004/Weak	-0.012/Weak	-0.015/Weak
education-num	0.033/Weak	-0.044/Weak	1.000/Perfect	0.121/Weak	0.078/Weak	0.146/Weak
capital-gain	0.074/Weak	0.004/Weak	0.121/Weak	1.000/Perfect	-0.031/Weak	0.080/Weak
capital-loss	0.056/Weak	-0.012/Weak	0.078/Weak	-0.031/Weak	1.000/Perfect	0.055/Weak
hours-per-week	0.069/Weak	-0.015/Weak	0.146/Weak	0.080/Weak	0.055/Weak	1.000/Perfect

2. For education, construct a bar chart of the variable, with an overlay of the target variable (income). Normalize the bar chart. Explain the results

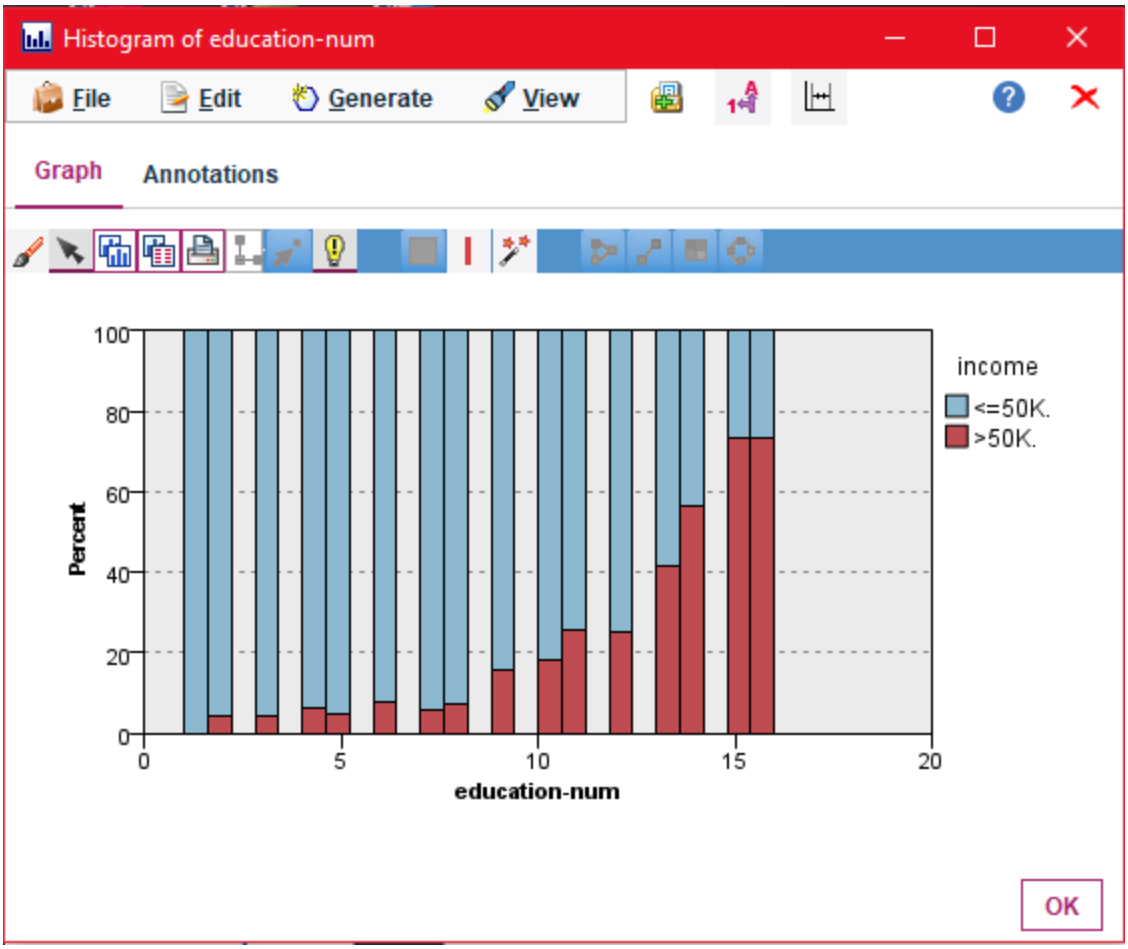
A simple histogram graph is developed for this and an overlay of target variable is then applied on the bar chart as shown below.



Overlaying the two attributes gives the following bar chart where it is difficult to notice the education-income relation



So the overlay is normalized to get the bar chart as below.

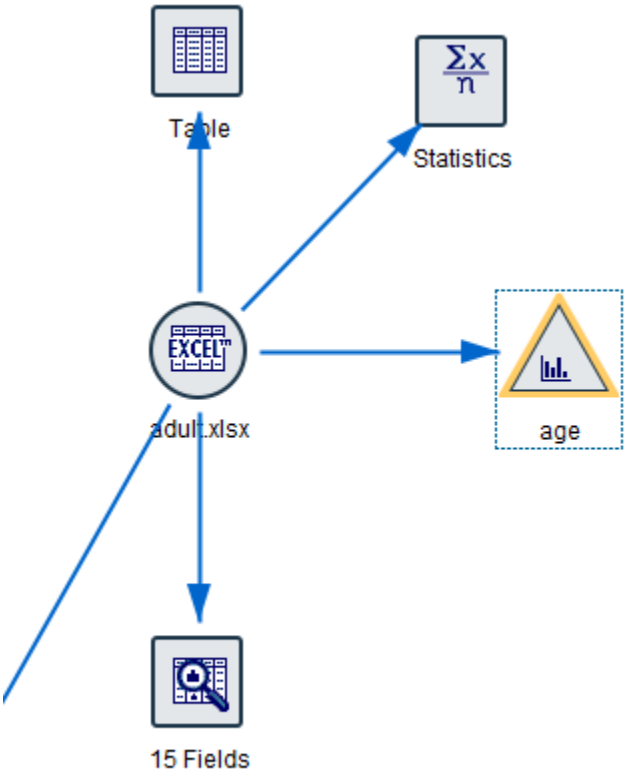


Now the bar chart clearly shows that income grows higher with education. For less grade in education gives the income less than 50K and people with higher education have an income greater than 50K. Moreover, a consistent higher income is seen in education number of 15 and the income less than 50K increases with the decrease in the education numbers.

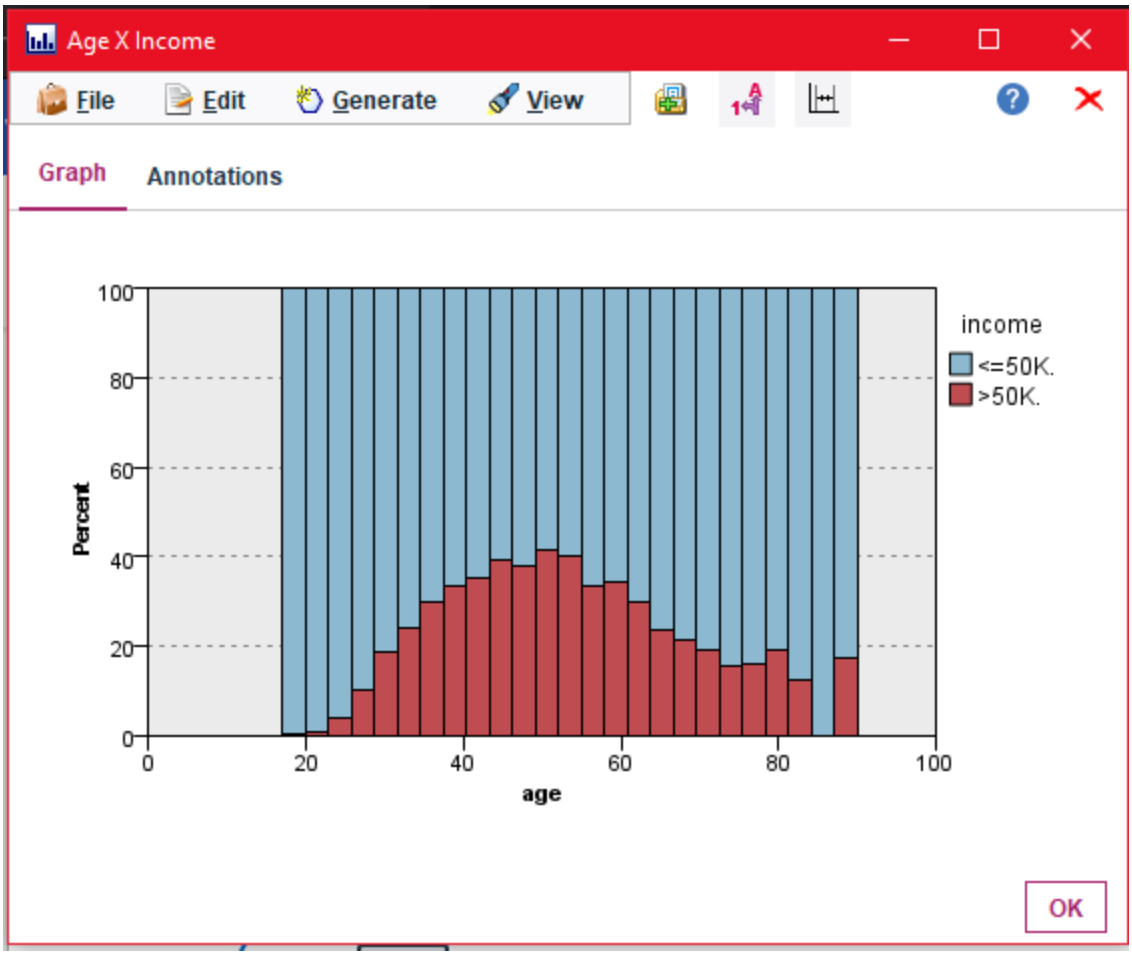
3. Construct a normalized histogram of age, with an overlay of the target variable income.

Explain the results

The attribute age is overlayed by the target income in the histogram of graph output as shown



When the result is normalized, the result is as below, shows that there is a higher level of income greater than 50K among the people with age between 40 and 60. Below 40 years of age, the higher level of income decreases and so happens beyond 60 years of age. Also, there is a slight increase in the income greater than 50K among the age of 80 and is completely less than 50K at 82 years. However, an abrupt increase of income is seen at the age of people 84-88 years of age.


Age X Income
File
Edit
Generate
View
Graph
Annotations
100
80
60
40
20
0
Percent
0
20
40
60
80
100
age
income
<=50K.
>50K.
OK