**Income Classification Using SAS: A Case Study with 1994 Census Data**

Nandan Patel

Contents

[**I.** **Dataset and task description** 3](#_Toc132395719)

[**II.** **Loading data** 5](#_Toc132395727)

[**III.** **Dataset characteristics** 5](#_Toc132395728)

[**IV.** **Examining categorical variables** 6](#_Toc132395732)

[**IV.1.** **Check and Correct Errors when necessary** 6](#_Toc132395733)

[**IV.2.** **Check for missing value.** 6](#_Toc132395734)

[**IV.3.** **Check for missing value (Informat Method)** 9](#_Toc132395735)

[**IV.4.** **Treat missing value** 10](#_Toc132395736)

[**IV.5.** **Create one or more derived variables.** 14](#_Toc132395737)

[**IV.6.** **Combine values in one for a categorical variable.** 14](#_Toc132395739)

[**V.** **Examining numerical variables** 16](#_Toc132395740)

[**V.1.** **Check Errors of Numerical Data** 16](#_Toc132395741)

[**V.2.** **Correcting Data:** 21](#_Toc132395742)

[**V.3.** **Detect Outliers:** 21](#_Toc132395744)

[**V.4.** **Remove Outliers:** 25](#_Toc132395745)

[**V.5.** **Test for normality** 27](#_Toc132395746)

[**V.6.** **Apply a transformation:** 29](#_Toc132395747)

### **Dataset and task description**

### **About The Dataset**

### This dataset has been taken from the famous UCI Machine Learning Repository. The data in the dataset was extracted from the 1994 Census Bureau. The prediction objective is to determine whether an individual earns more than $50,000 annually.

**Description**

The dataset include 15 variables, including:

* 9 characteristic variables:
  + Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
  + education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
  + MaritalStatus: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
  + Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
  + Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
  + Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
  + Sex: Female, Male.
  + NativeCountry: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
  + Income: >50K,<=50K
* 6 numerical variables:
  + Age: Describes the age of individuals. Continuous.
  + fnlwgt: Continuous.
  + Educationnum: Number of years spent in education. Continuous.
  + Capitalgain: Continuous.
  + Capitalloss: Continuous.
  + Hoursperweek: Continuous.

### The dependent variable is “Income”, and the independent variables in the dataset are the rest except for fnlwgt (which was used to indicate the number of times or number of similar instances appeared in the dataset). These independent variables are used to predict the dependent variable, which is income. Income is either above or below $50,000 and is categorized as ">50K" or "<=50K".

### The project uses SAS software to perform data cleaning techniques learnt throughout the course to check and correct errors in data and prepare data for analysis. There are various tasks associated with the income dataset, including:

### Importing the dataset: The first step is to import the income dataset into SAS and ensure that it is correctly formatted and structured.

### Data cleaning: The dataset may contain missing values, unstandardized or errored inputs that need to be adjusted before analysis.

### Correcting errors: perform techniques to standardize the data.

* Examining characteristic and numerical variables to correct if there are any unnormal characteristic of the dataset (outliers, data distribution, etc.), ensuring the quality of outcomes when doing analysis.

### **Loading data**

Importing the income dataset into SAS and ensuring that it is accurately formatted and structured is the initial step.

**Code in SAS:**

Graphical user interface, text, application

Description automatically generated

**The result:**

A screenshot of a computer

Description automatically generated with medium confidence

### **Dataset characteristics**

### The target variable "Income" is a binary variable with two values of:

### "<=50K" accounted for 75.92%

### ">50K" accounted for 24.08%;

**Code in SAS:**

A picture containing text

Description automatically generated

**The results:**

Table

Description automatically generated

### **Examining categorical variables**

### **Check and Correct Errors when necessary**

Based on the results, there are 9 Categorical variables, which are all formatted as characters. They are Education, Income, MartialStatus, NativeCountry,Occupation, Race, Relationship, Sex, Workclass;

**Code in SAS:**



**The Results**

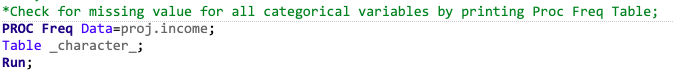
Table

Description automatically generated

### **Check for missing value.**

Based on the frequency table, 3 out of 9 categorical variables have missing values, which are workclass (1836 missing, 5.64% of total observations), occupation (1843missing, 5.66 of total observations) and nativecountry (583 missing, 1.79% of total observations)

**Code in SAS**



Table

Description automatically generated**The Results**

Table

Description automatically generatedTable

Description automatically generated with medium confidence

### **Check for missing value (Informat Method)**

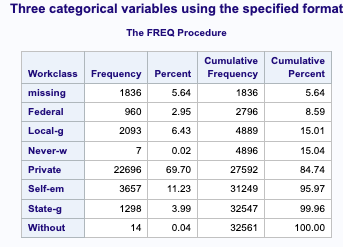
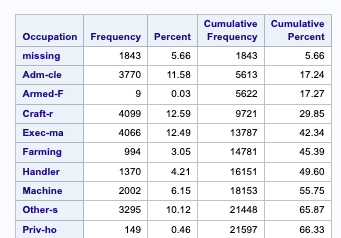
We also use Informat Method to check for the missing value with “?” from the three variables.

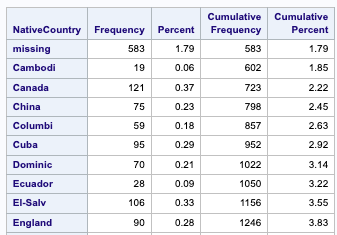
**Code in SAS**

**Text

Description automatically generated**

**The Results**

**** ****

****

### **Treat missing value**

**Code in SAS**

Text

Description automatically generated with low confidence

**The Results**

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedGraphical user interface

Description automatically generated

Based on the results obtained from the analysis, it was decided to delete all missing values:

* Reason 1 the proportion of missing values is small (less than 5% or so), you consider deleting the observations with missing values"?"
* Reason 2 Consider the impact of missing values: the missing values are not likely to bias the results or affect the interpretation of our analysis, deleting the missing values may be more appropriate.

**Code in SAS:**

Text, table

Description automatically generated

**The Results**

Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated with medium confidence

**Code in SAS:**

Graphical user interface, application

Description automatically generated

**The Results**

Table

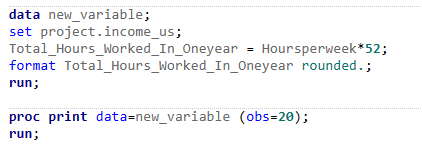
Description automatically generated

### **Create one or more derived variables.**

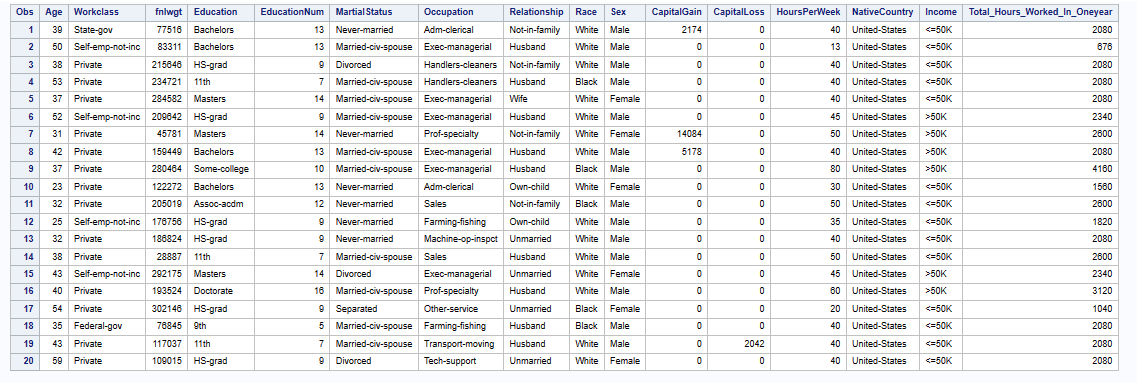
### The data step creates a new variable named "Total\_Hours\_Worked\_In\_Oneyear" by multiplying the existing variable "Hoursperweek" by 52, which represents the number of weeks in a year.

* The new variable is formatted to display rounded values using the "rounded." format in the "format" statement. This ensures that the resulting values are integers instead of decimals.
* The "set" statement in the first data step reads in the "project.income" dataset and assigns it to a new SAS dataset called "new\_variable".
* The "run**" statement ends the data step.**

**Code in SAS**

****

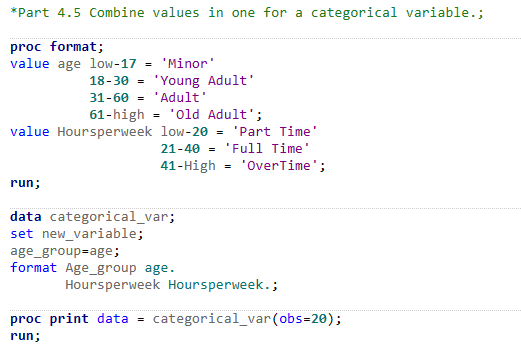
**The Result**

****

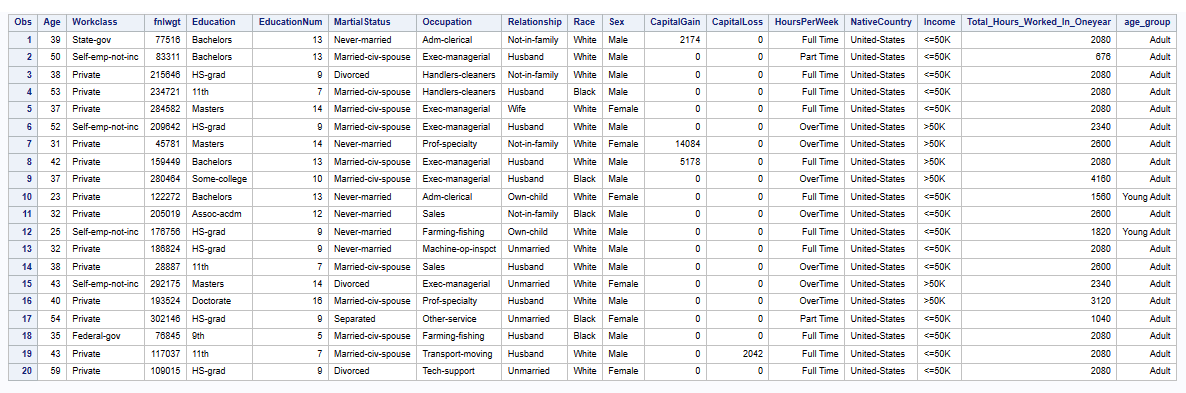
### **Combine values in one for a categorical variable.**

* The data step creates a new dataset named "categorical\_var" by copying the contents of the "new\_variable" dataset created in the first data step.
* Two user-defined formats are applied to the "Age" and "Hoursperweek" variables in the "format" statement. The "age." format categorizes the "Age" variable into four groups based on age range, while the "Hoursperweek." format categorizes the "Hoursperweek" variable into three groups based on the number of hours worked per week.
* The "proc print" statement is used to print the resulting "categorical\_var" dataset.

**Code in SAS**

****

**The Result**



### **Examining numerical variables**

### **Check Errors of Numerical Data**

Use PROC CONTENT to check on variables of dataset, Identifying missing or incorrect variable types or formats.

**Code in SAS**

Graphical user interface, text, application

Description automatically generated

**The results:**

* + - There are 6 Numeric variables, which are all formatted as numeric.
    - Age, CapitalGain, CapitalLoss, EducationNum, HoursPerWeek, fnlwgt

Table

Description automatically generated

Use PROC MEANS to examine numeric variables and searching for errors if any.

**Code in SAS:**Used N for total number of observations, nmiss to check missing values, mean for calculating average, mode for the central tendency, stddev for Standard Deviation with Minimum and Maximum, and maxdec to check on accuracy of data.

A picture containing text

Description automatically generated**The results:**

* + - It’s clearly that there are no missing values across numeric variables.
    - Checking on other values, its noticeable that there maybe outliers in variables:
      * Max value of Age is 90 and Min Value is 17, people whose age range from below 18 and > 80 rarely earn income.
      * Max value of HoursPerWeek is 99, means that some people work 14 hours per week.
      * Min of EducationNum is 1, means that people just have 1 year of education, is also rare cases.
    - Fnlwgt (Final Weight) indicates the number of other observations share the same value of that observation. We won’t use this variable in this case.
    - Mean and Mode value of Capital variables present values of 0, check on Frequency of these variable to investigate further.

***Table

Description automatically generated***

Use PROC MEANS to examine proportion of numeric variables.

**Code in SAS:**

Graphical user interface, text, application, email

Description automatically generated

**The results:**

* + - 0 values contribute more than 90% of Capital Gain and Capital Loss variables.
    - These two variables would be dropped as it could not be used for analysis.
    - Other variables show no further errors.

Table

Description automatically generated Table

Description automatically generated

Use PROC UNIVARIATE to examine the rest numeric variables with options to check histogram distribution and listing Extreme Values

**Code in SAS:**

**Graphical user interface, text, application

Description automatically generated**

**The results:**

* There are possible extreme values in these variable impacted on the distribution of the values

Table

Description automatically generated**Table

Description automatically generatedTable

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

### **Correcting Data:**

### Drop variables Fnlwgt, CapitalGain and CapitalLoss

**Code in SAS:**

Graphical user interface, text, application, email

Description automatically generated

**The results:**

**Table

Description automatically generated**

### **Detect Outliers:**

Detect outliers based on Histogram and box plot.

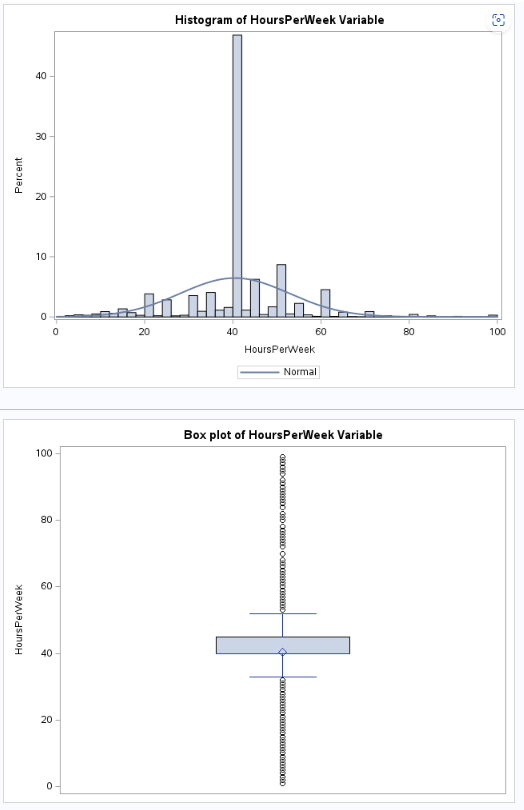
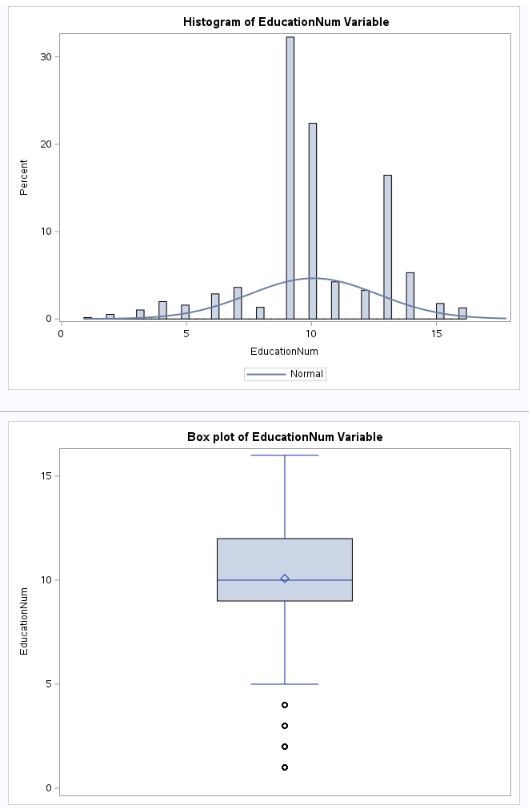
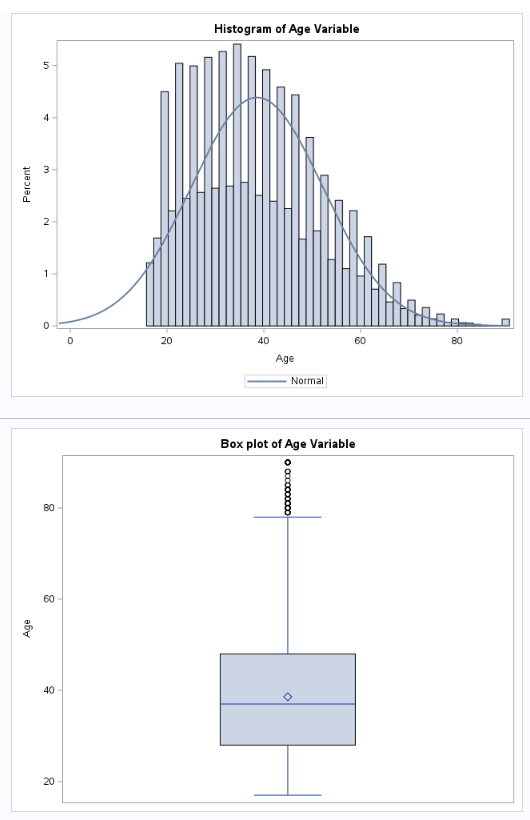
**Code in SAS*:*** using PROC SGLOT to visualize histogram and Box Plot for the variables

**Text

Description automatically generated**

**The results:**

* + - The variables have normal distribution.
    - There are outliers detected in box plot of three variables.



Listing outliers.

**Code in SAS:**

* Detect outliers for Age and HoursPerWeek variable using Standard Deviation method considering they have normal distribution
* Using Interquartile range to detect outliers for EducationNum variable
* Then, listing the outliers based on the defined range of mentioned methods

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**The results:**

A picture containing text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

### **Remove Outliers:**

Remove the outliers.

**Code in SAS:**

* + For Age and HoursPerWeek, the Standard Deviation was used to define the range to remove outliers.
  + For EducationNum, the Interquartile was used to define the range to remove outliers.
  + Histogram and Box Plot adapted to examine variable after removing outliers.

**Graphical user interface, text, application

Description automatically generated**

Text

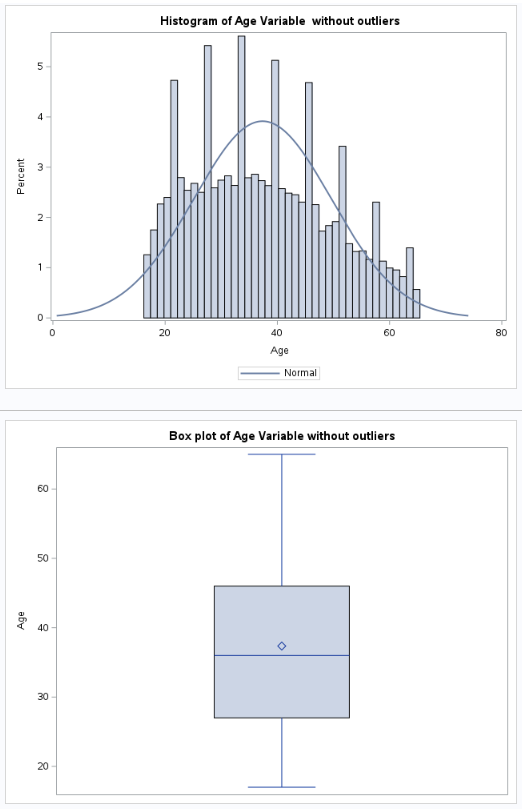
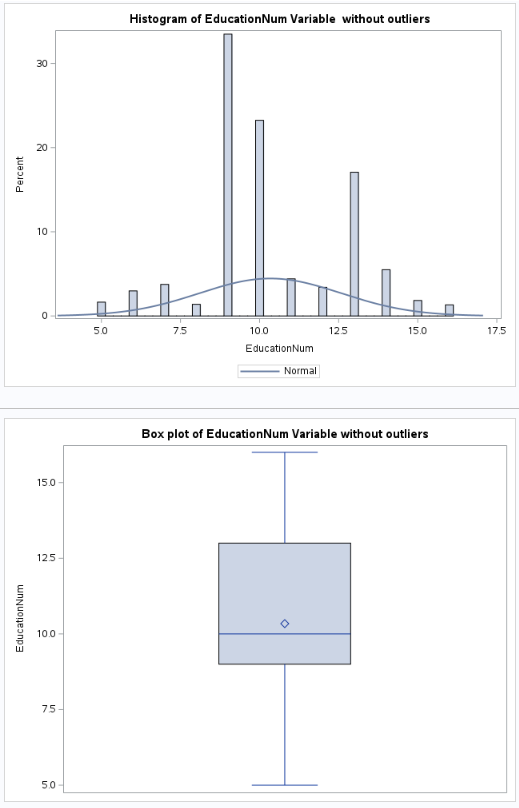
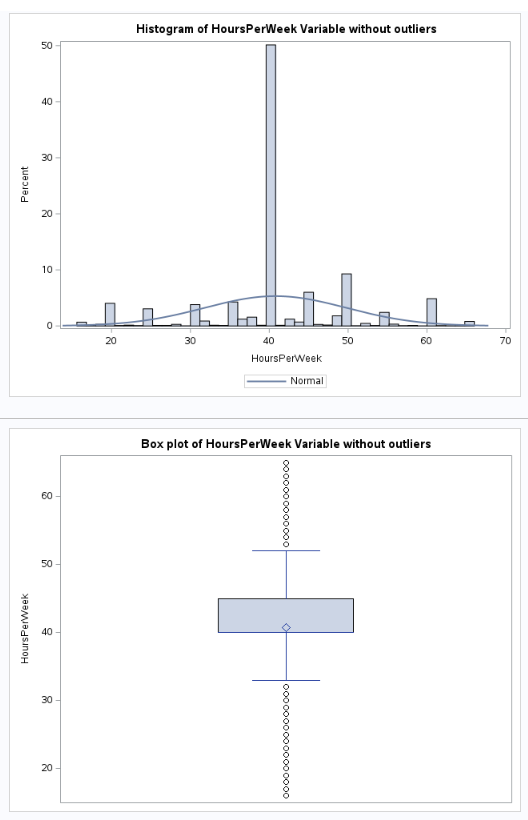
Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**The results:**

* + - After removing the outliers, normal distribution could be observe across the variables
    - Note that there are some outliers remain in HoursPerWeek but woud not impact on the dataset.



### **Test for normality and plot histogram and QQ plots for "Age" variable with a skewed distribution**

**Code in SAS:**

**Graphical user interface

Description automatically generated with medium confidence**

**The result:**

Based on the result of the code, we can detect "Age" variable is not normally distributed:

* + Histogram: the histogram is right-skewed
  + QQ plot: the points deviate from the line.
  + Normality tests: the p-value for the normality test KS resulted in a p-value of less than 0.01, which suggests that the variable is significantly different from a normal distribution.

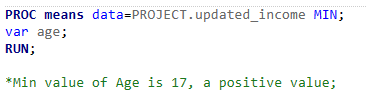
**Chart

Description automatically generated** **Chart, histogram

Description automatically generated**

### **Apply a transformation:**

This step to transform the Age variable to normal distribution:

**Code in SAS:** check again the min value of Age to apply Log transformation:

**The result:** the min value of Age is 17, a positive value.

**Code in SAS:** apply Log transformation and test again the normality:

Graphical user interface, text, application, email

Description automatically generated

**The result:**

A screenshot of a computer

Description automatically generated with medium confidence

**Graphical user interface, chart, histogram

Description automatically generated**