We begin by importing the necessary libraries, initiating a Spark session, and reading in the data. Here, we can see what the data looks like:

Graphical user interface

Description automatically generated with medium confidence

The first data preprocessing step is to rename all columns that contain a “.”, because this character will cause problems later. We will replace it with an underscore:

Text

Description automatically generated

Next, we will encode the Education column into a new integer column. Since levels of education have an implied order, we can provide additional meaning to this column by encoding it into an ordinal variable:

A picture containing text

Description automatically generated

After doing this, we will look at the various columns, based on their data type. First, we will look at the categorical/string columns:

Graphical user interface, text, application

Description automatically generated

We also will examine the numerical columns and their statistics:

Graphical user interface, text, application

Description automatically generated

Text, application

Description automatically generatedChart, bar chart

Description automatically generatedText

Description automatically generatedSchematic

Description automatically generated with low confidenceGraphical user interface, text

Description automatically generatedChart

Description automatically generatedTable

Description automatically generatedText

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedText

Description automatically generatedChart, histogram

Description automatically generatedGraphical user interface, text, application

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generatedText

Description automatically generatedChart, bar chart

Description automatically generated