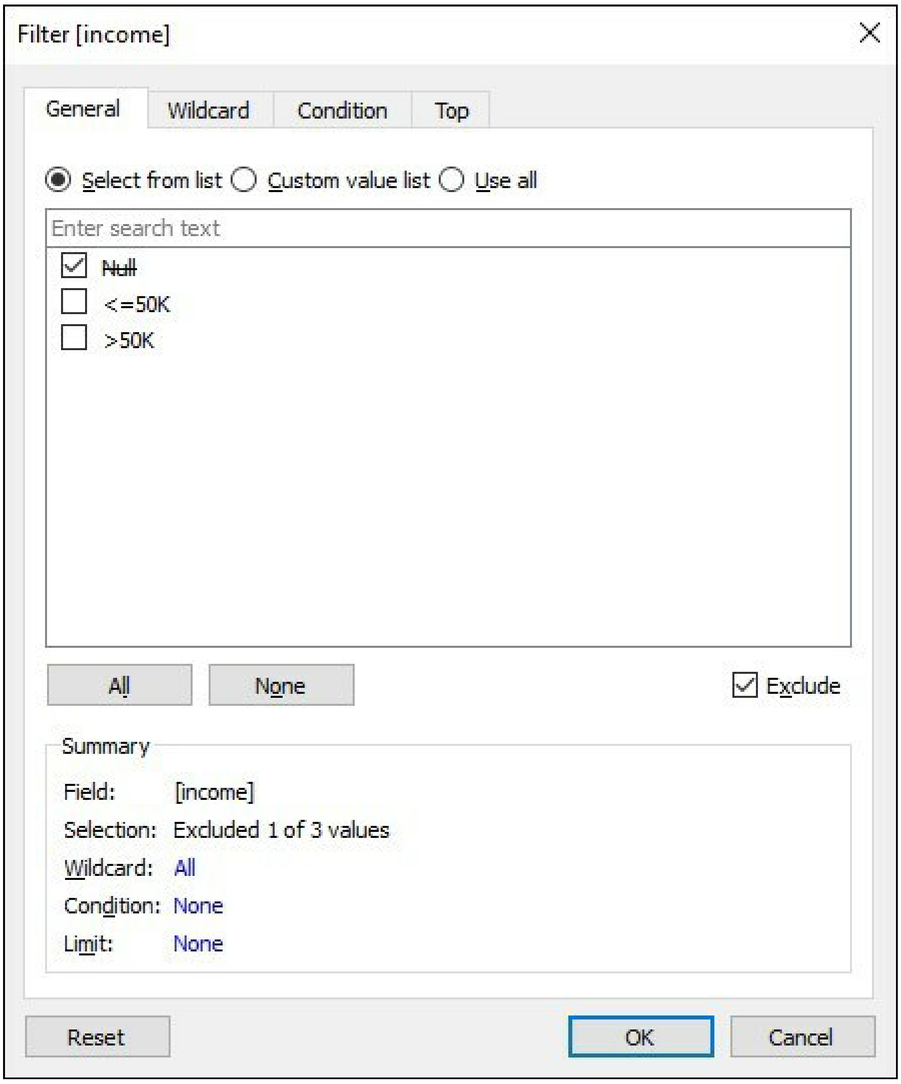
The dataset used in this project has 49,000 records. You can see from the files that the data has been divided into a training dataset and a test set. The training dataset contains approximately 32,000 records and the test dataset around 16,000 records. It's helpful to note that there is a column that indicates the salary level or whether it is greater than or less than fifty thousand dollars per annum. This can be called a binomial label, which basically means that it can hold one or two possible values. When we import the data, we can filter for records where no income is specified. There is one record that has a NULL, and we can exclude it. Here is the filter:



Filter [income]

● Select from list ○ Custom value list ○ Use all

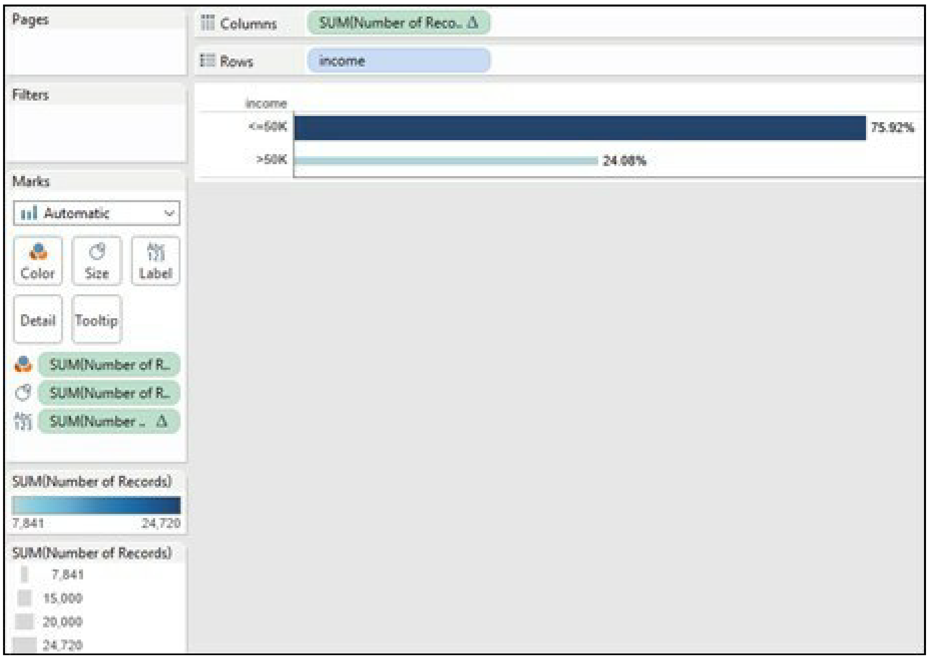
☑ ~~Null~~
☐ <=50K
☐ >50K

☑ Exclude

Summary
Field: [income]
Selection: Excluded 1 of 3 values
Wildcard: All
Condition: None
Limit: None

Let's explore the binomial label in more detail. How many records belong to each label?

Let's visualize the finding. Quickly, we can see that 76 percent of the records in the dataset have a class label of <50K.

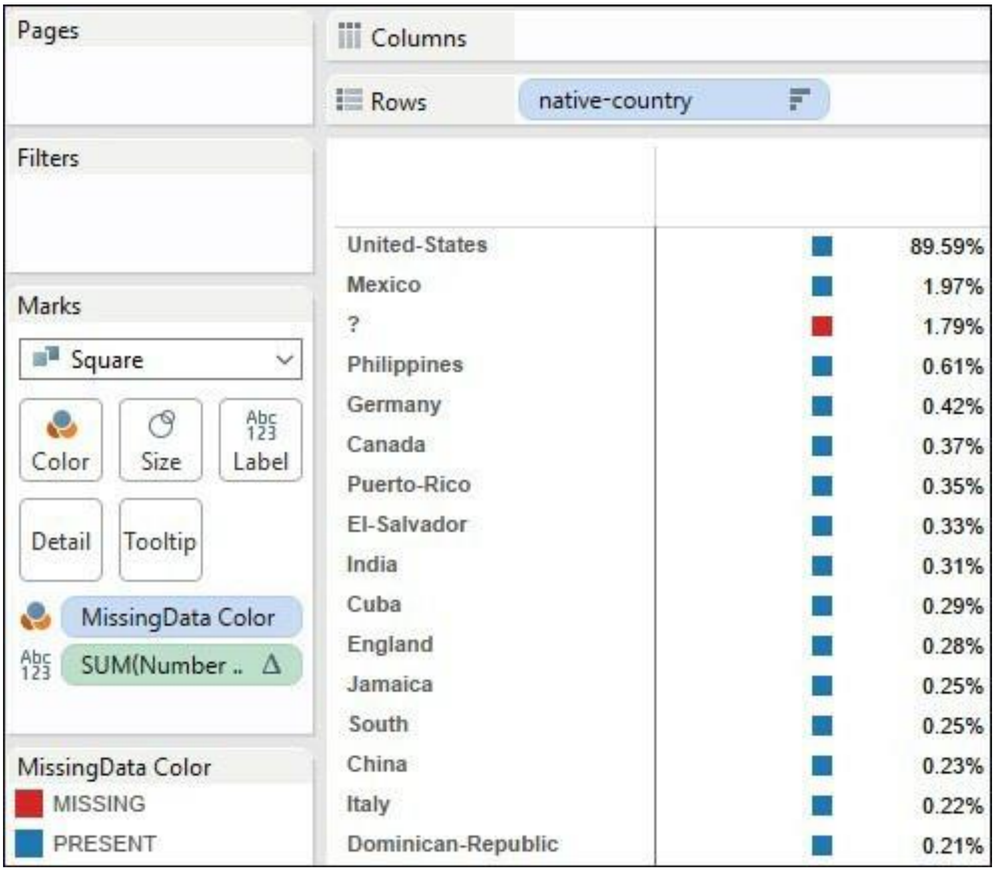


Let's have a browse of the data in Tableau in order to see what the data looks like. From the grid, it's easy to see that there are 14 attributes in total. We can see the characteristics of the data:

Seven polynomials: workclass, education, marital-status, occupation, relationship, race, sex, native-country

One binomial: sex

Six continuous attributes: age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week



Rows: native-country

Marks: Square

MissingData Color

SUM(Number .. Δ

United-States	🟦	89.59%
Mexico	🟦	1.97%
?	🟥	1.79%
Philippines	🟦	0.61%
Germany	🟦	0.42%
Canada	🟦	0.37%
Puerto-Rico	🟦	0.35%
El-Salvador	🟦	0.33%
India	🟦	0.31%
Cuba	🟦	0.29%
England	🟦	0.28%
Jamaica	🟦	0.25%
South	🟦	0.25%
China	🟦	0.23%
Italy	🟦	0.22%
Dominican-Republic	🟦	0.21%

MissingData Color

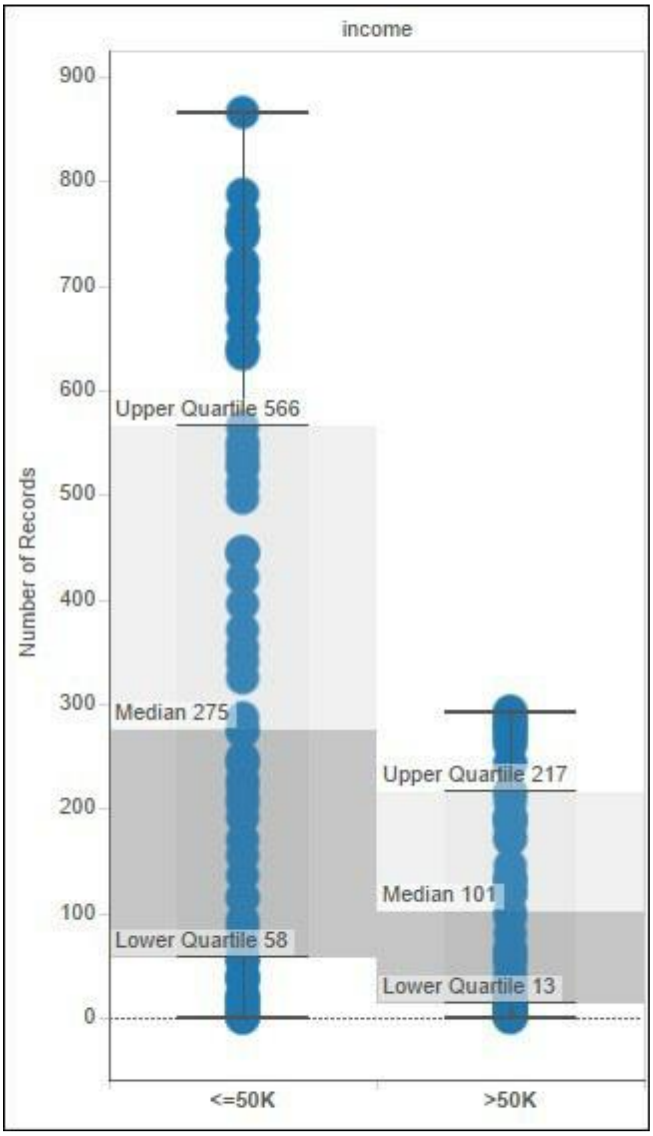
🟥 MISSING

🟦 PRESENT

From the preceding chart, we can see that nearly 2 percent of the records are missing for one country, and the vast majority of individuals are from the United States. This means that we could consider the native-country feature as a candidate for removal from the model creation, because the lack of variation means that it isn't going to add anything interesting to the analysis.

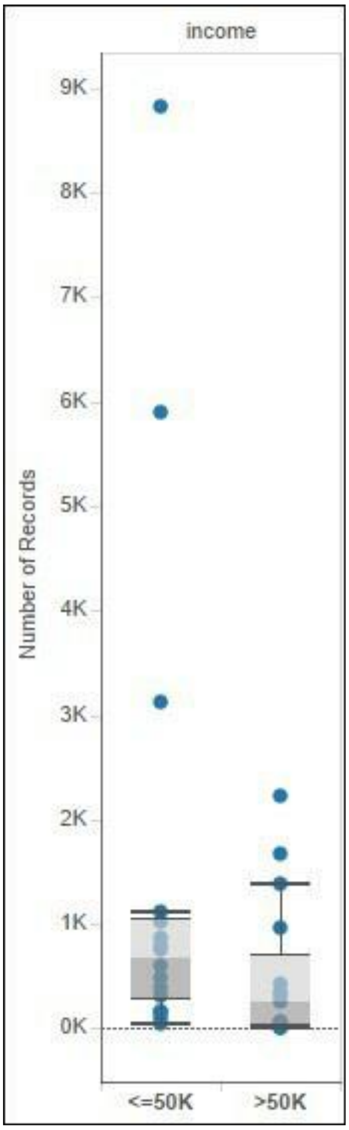
Data exploration

We can now visualize the data in boxplots, so we can see the range of the data. In the first example, let's look at the age column, visualized as a boxplot in Tableau:



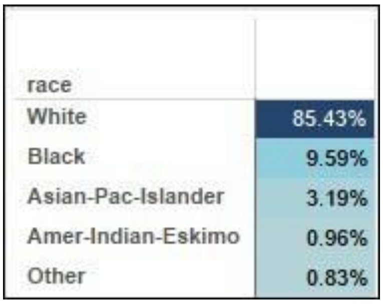
We can see that the values are higher for the age characteristic, and there is a different pattern for each income level.

When we look at education, we can also see a difference between the two groups:



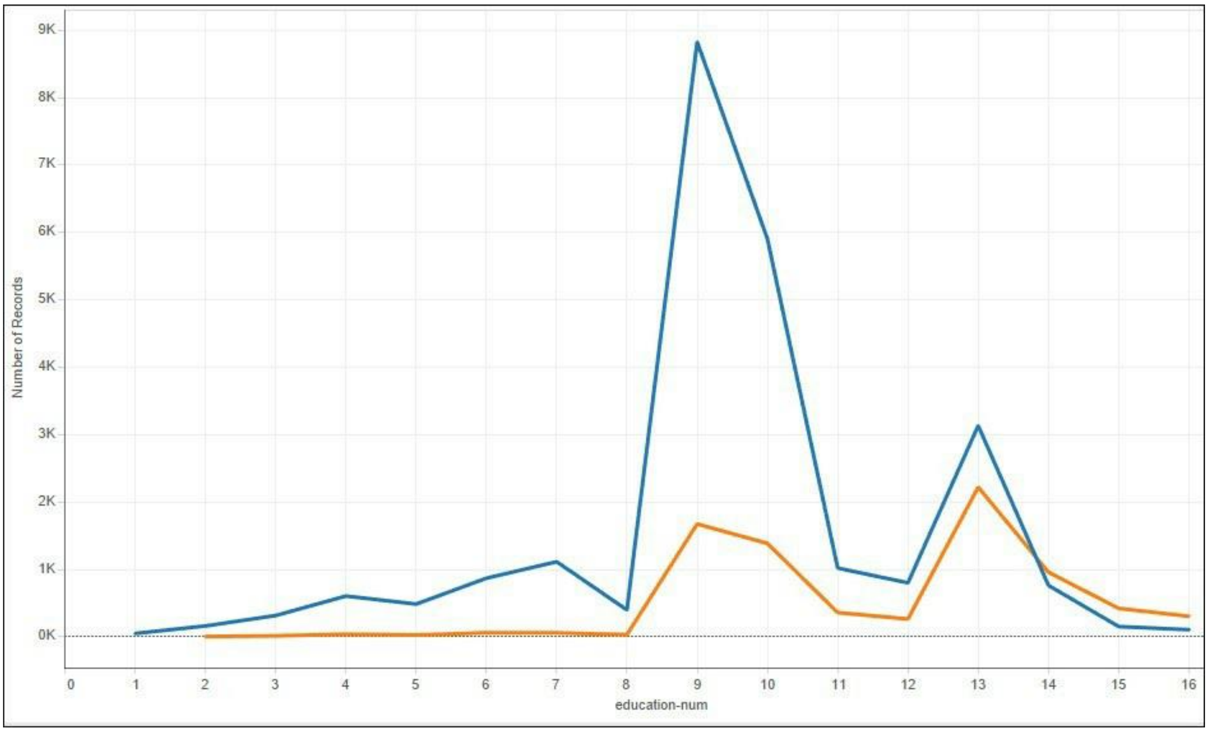
We can focus on age and education, while discarding other attributes that do not add value, such as native-country. The fnlwgt column does not add value because it is specific to the census collection process.

When we visualize the race feature, it's noted that the White value appears for 85 percent of overall cases. This means that it is not likely to add much value to the predictor:



race	
White	85.43%
Black	9.59%
Asian-Pac-Islander	3.19%
Amer-Indian-Eskimo	0.96%
Other	0.83%

Now, we can look at the number of years that people spend in education. When the education number attribute was plotted, then it can be seen that the lower values tend to predominate in the <50K class and the higher levels of time spent in education are higher in the >50K class. We can see this finding in the following figure:



This finding may indicate some predictive capability in the education feature. The visualization suggests that there is a difference between both groups since the group that earns over $50K per annum does not appear much in the lower education levels.

To summarize, we will focus on age and education as providing some predictive capability in determining the income level.

The purpose of the model is to classify people by their earning level. Now that we have visualized the data in Tableau, we can use this information in order to model and analyze the data in R to produce the model.