1 Properties of the Dataset

The dataset has no missing attribute values, which avoids the challenge of having to fill in or compensate for any missing values. In lieu of this, practically all numerical attributes in the dataset are unimodally distributed with a positive skew in the range of 1 to 7.5. For instance, a histogram of the Product Related attribute values seen in fig. 1.1 shows the positively skewed distribution of the attribute.

Figure 1.1: Histogram of the Product Related attribute values.

The significance of the distribution increases for the target attribute we want to predict. The Revenue attribute has two classes, TRUE and FALSE, representing whether a visitor eventually bought an item from Nozama's website. Of the 12 330 records in the dataset, 84.5 % of the values for Revenue are FALSE, which implies that any predictor could easily attain a prediction accuracy of 84.5 % if it always predicted Revenue to be false. Therefore, the evaluation of predictors should factor this in as a minimum prediction accuracy to achieve.

Given that both the prediction attributes and the target attribute have similar distributions, the correlation between the attributes were observed next. From fig. 1.2, two relatively significant correlations for Exit Rates and Page Values can be observed in relation to revenue. One should expect to see that Exit Rates and Page Values play in important role in predicting the target attributes, which will be explored further in later steps of the analysis.

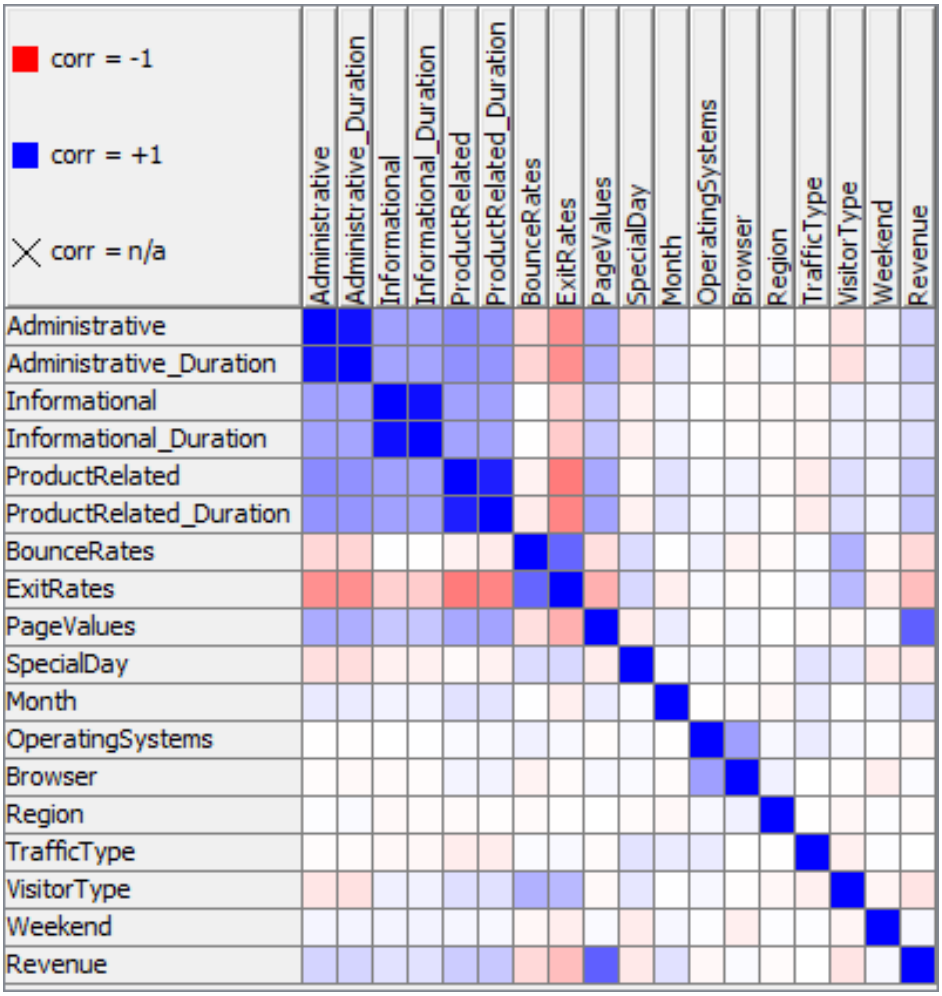

Figure 1.2: Rank correlation matrix of the attributes.

As far as why the data is distributed this way, one could reason that a majority of visitors to Nozama's website tend to not interact much with it and don't make any purchases, leading to the high number of low values for each numerical attribute.