# Company data Random Forest Algorithm

Random Forest

Assignment

About the data:

Let’s consider a Company dataset with around 10 variables and 400 records.

The attributes are as follows:

 Sales -- Unit sales (in thousands) at each location

 Competitor Price -- Price charged by competitor at each location

 Income -- Community income level (in thousands of dollars)

 Advertising -- Local advertising budget for company at each location (in thousands of dollars)

 Population -- Population size in region (in thousands)

 Price -- Price company charges for car seats at each site

 Shelf Location at stores -- A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site

 Age -- Average age of the local population

 Education -- Education level at each location

 Urban -- A factor with levels No and Yes to indicate whether the store is in an urban or rural location

 US -- A factor with levels No and Yes to indicate whether the store is in the US or not

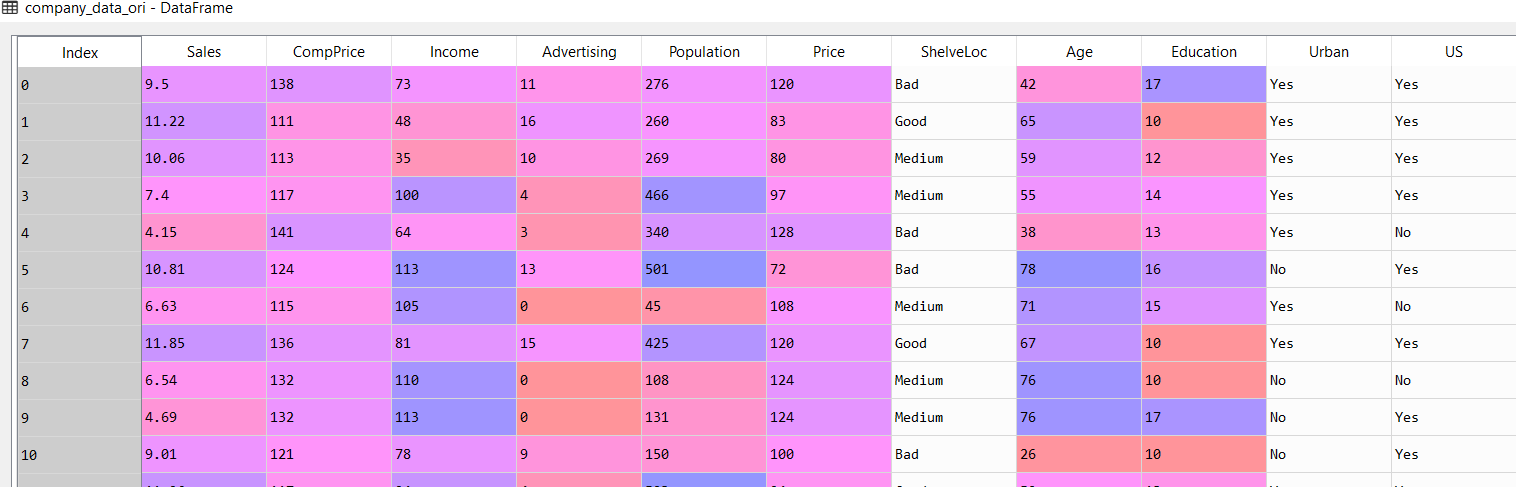
The company dataset looks like this:

Problem Statement:

A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

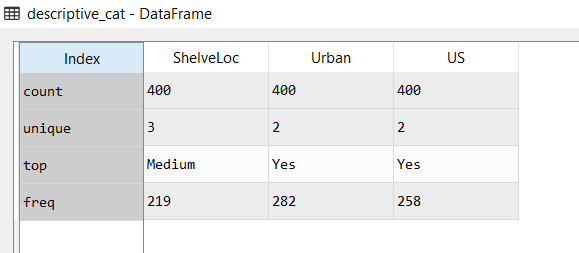
Approach - A Random Forest can be built with target variable Sales (we will first convert it in categorical variable) & all other variable will be independent in the analysis.

The following is the dataset

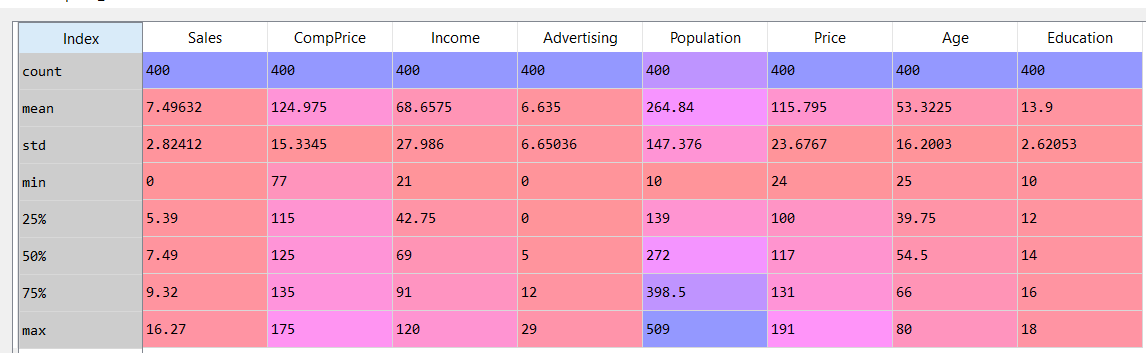


In the above data we have ShelveLoc, Urban and US are the categorical variables and the rest are continuous variables

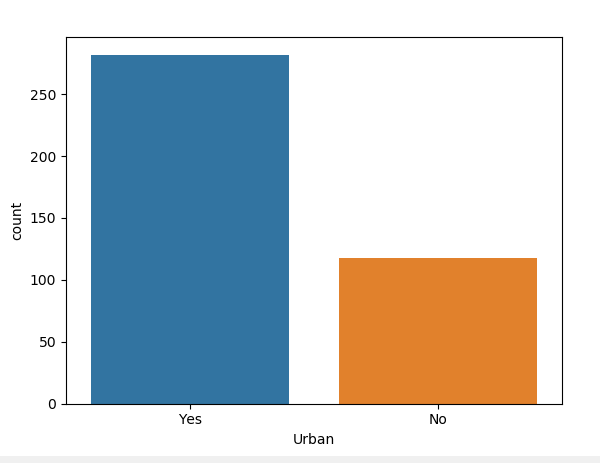
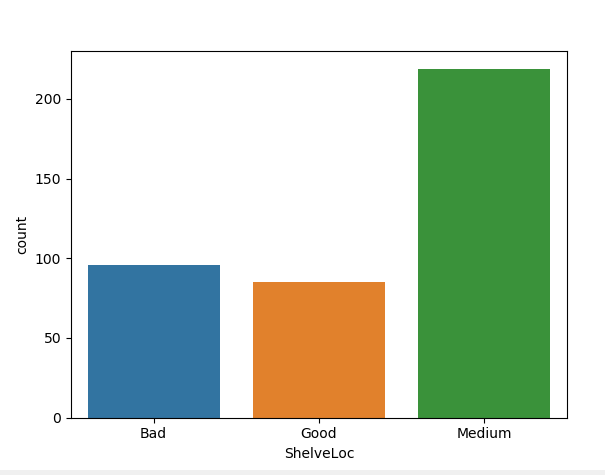
The following is the descriptive statistics of categorical variables:

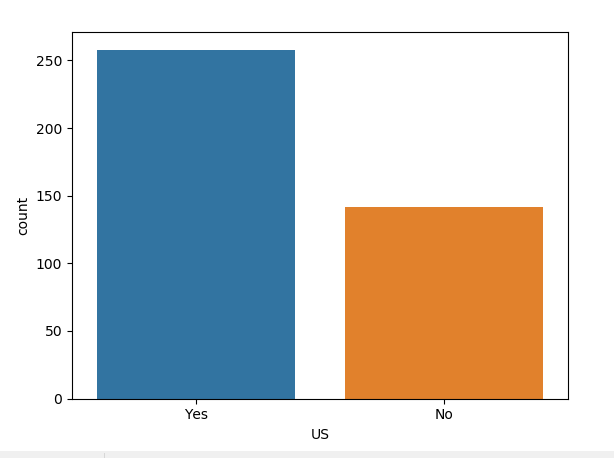


The following is the descriptive statistics of continuous variables:

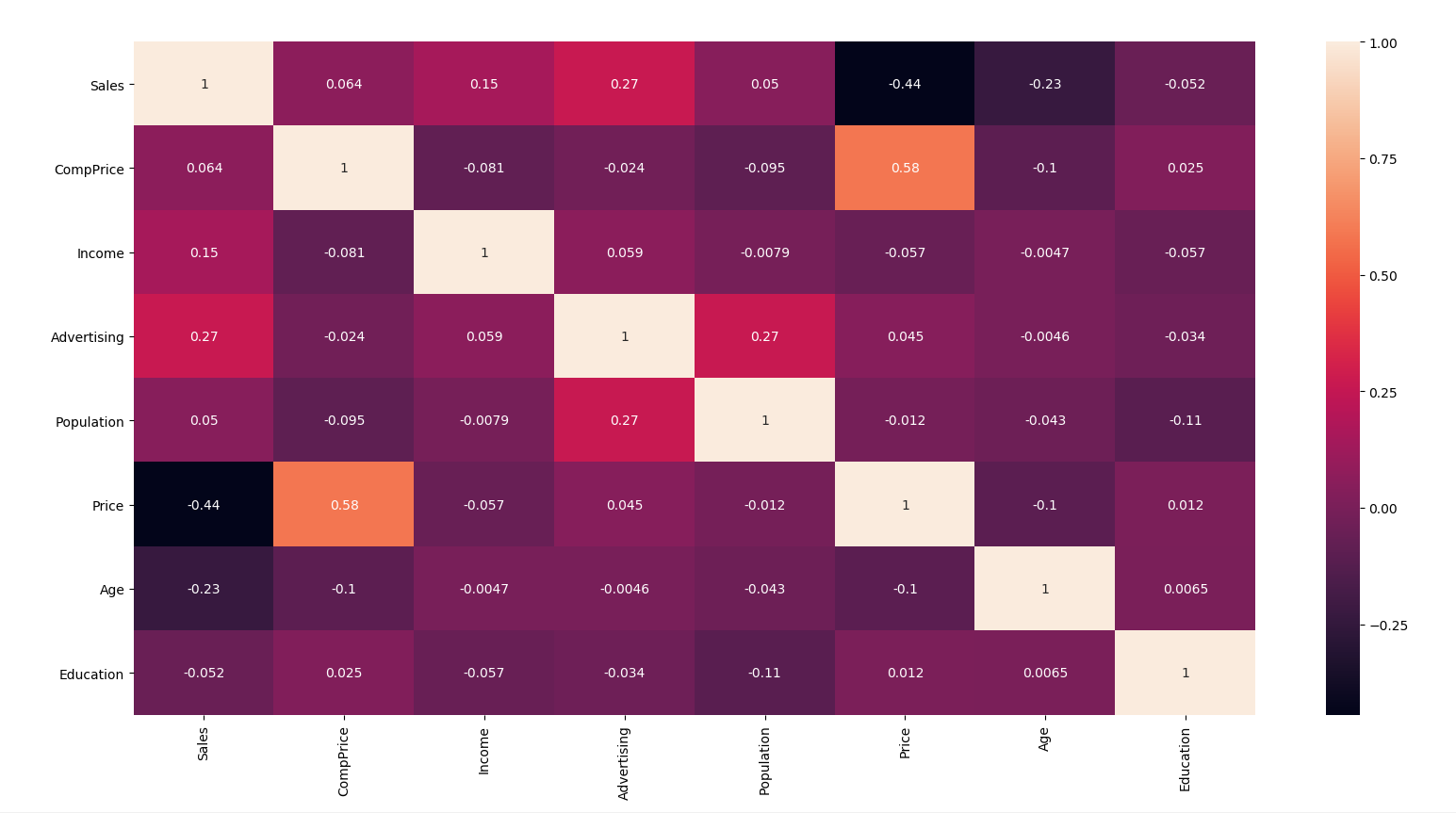


The following are the count plots of the continuous variables:



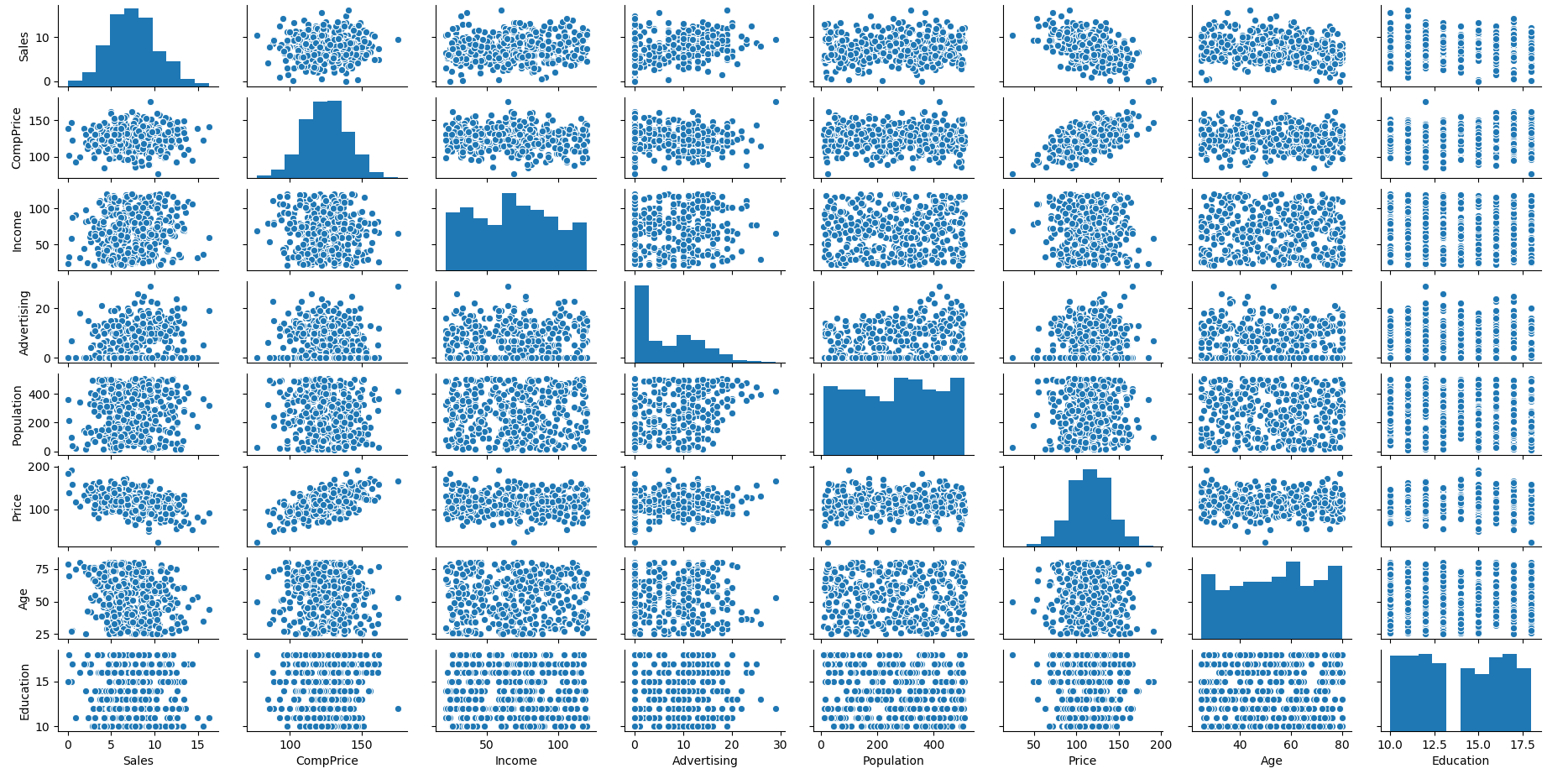


The following is the heatmap of the continuous variables of Company\_data

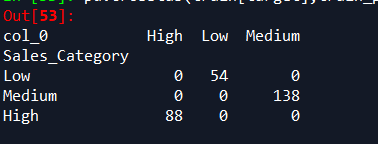


From the above matrix, all of the variables have low correlation

The following is the pair plot of the continuous variables/ data



We then pass the company training data to the random forest algorithm. The following are the results:



The results obtained are with an accuracy of 100%

We then pass the company test data to the algorithm and obtain the following results:

The results obtained are with an accuracy of 71%