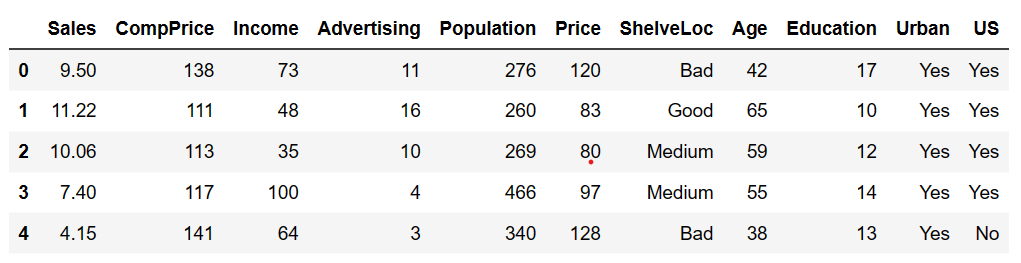
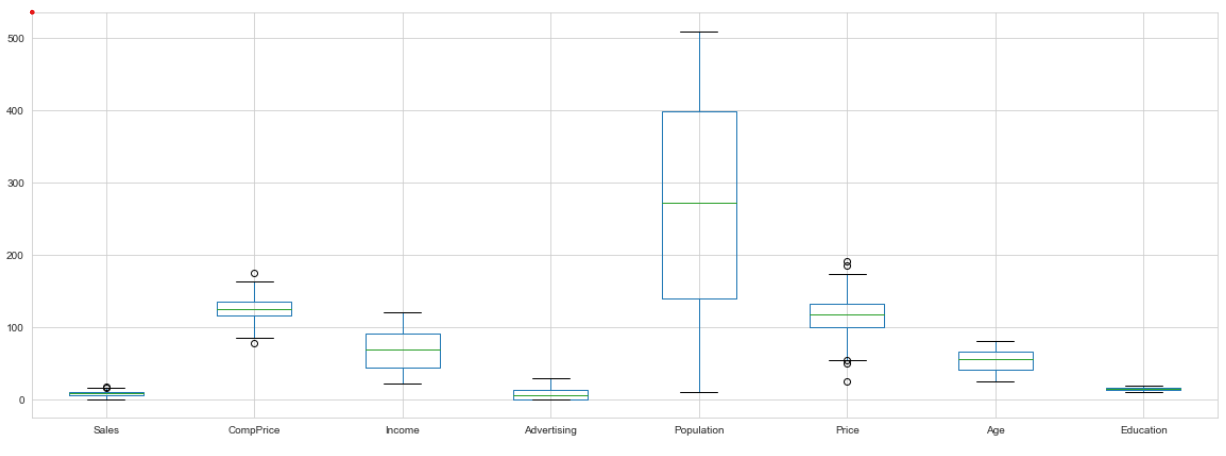
**Analysis - Company\_Data.csv**

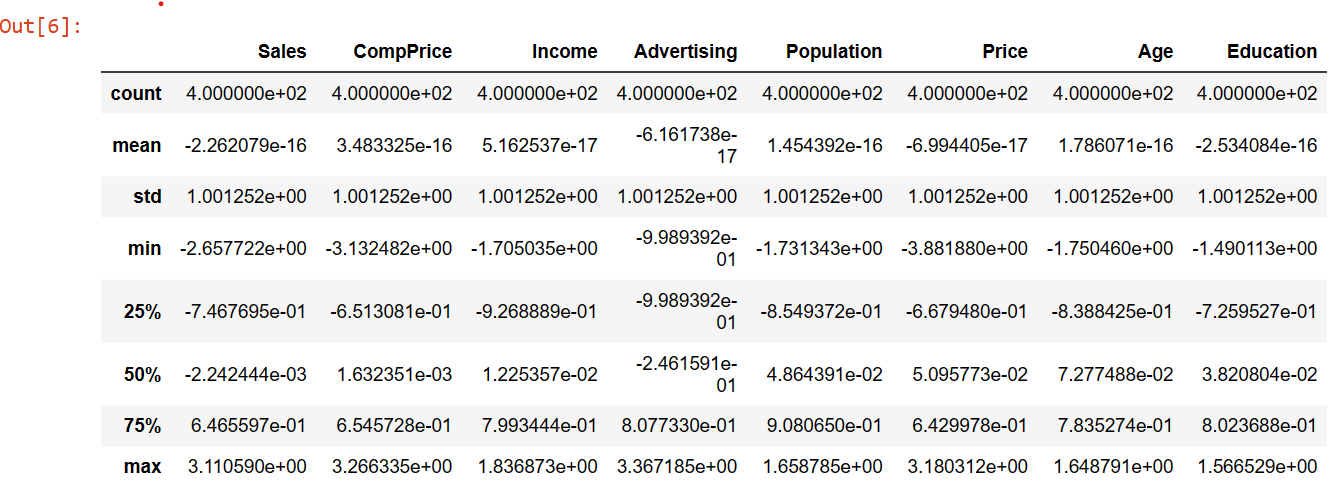
****

Company data has 10 columns, as shown in above table.

Data Analysis:

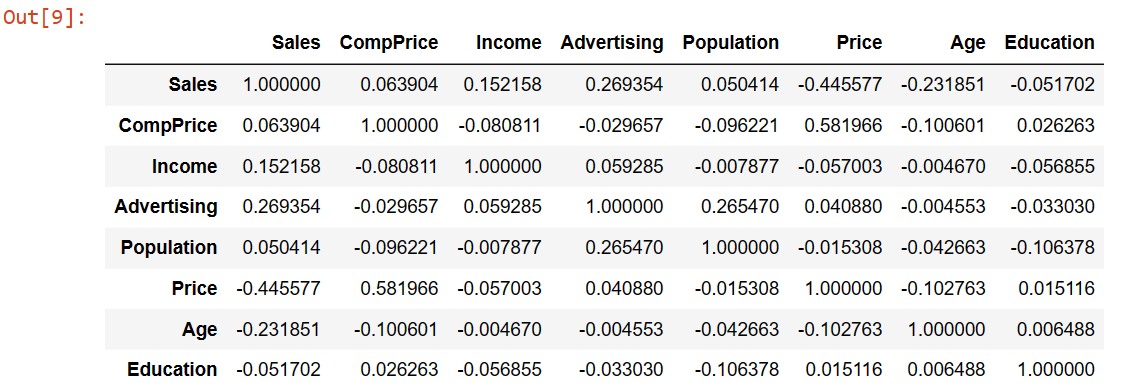
1. There is no missing value in any of the column.
2. Outlier Detection





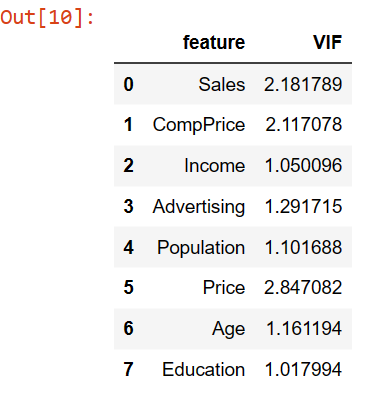
Boxplots for different features shows that there are few outliers in **Sales**, **CompPrice** & **Price** variables. Same can be confirmed via z-score details as only these three variables have z-scores beyond +-3, which means these variables have outliers.

1. Correlation Analysis:



Correlation table shows that no two variables have significant correlation. So, Collinearity does not exist.

1. Variance Inflation Factor (VIF): VIF tells that how much one feature explains the variability in data. VIF < 5 infers that the variable is not co-linear with any other feature.

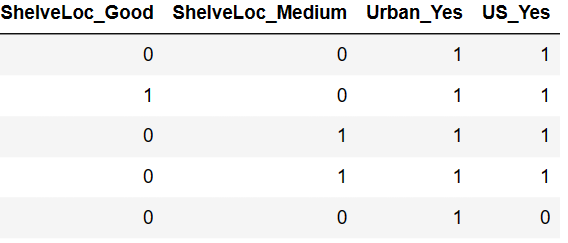


Since all VIF values are less than 5, suggests that multi-collinearity does not exist in the data.

Feature Engineering:

1. Data contains three categorical variables, which need to be converted to numeric values in order to feed to the model.

One hot encoding is used for the purpose, with one category dropped, just to avoid any ingested multi-collinearity.



1. Remove outliers- Any **value < mean – 3 std** has been replaced with **mean – 3 std**

Similarly, **values > mean + 3 std** has been replaced with **mean + 3 std**.

1. Dimensionality reduction is not required for two reasons

1- There are not too many features

2- We saw in correlation analysis that there is no significant correlation among features.

Proving the same point using PCA

As can be seen that first few transformed features are not explaining the almost all the variance in data, and even last transformed feature is significantly explaining the variance.

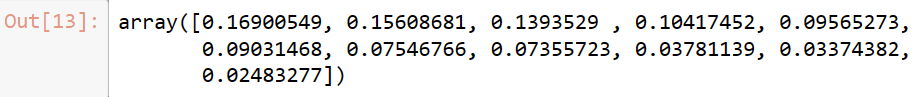
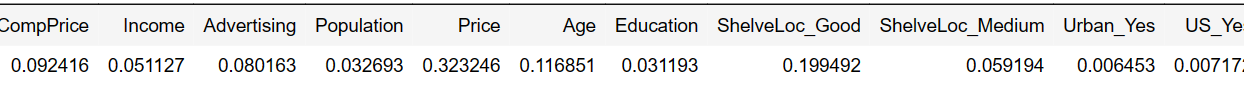


Image shows the explained variance by different transformed features, though I have taken number of PCA components as equal to the total number of features, even the last component has significant variance explanation.

So, dimensionality reduction does not make any sense to this data.

Feature Importance:

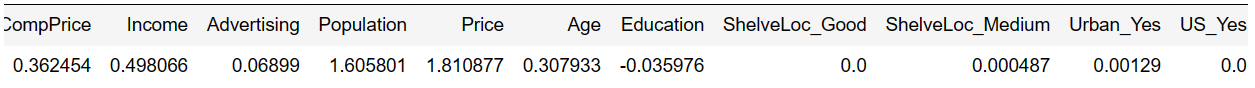
1. Feature Importance can be calculated based on multiple techniques.
2. Based on Random Forest
3. Permutation based
4. Random forest is a tree-based bagging technique, which trains different trees taking bootstrapped samples and different combination of features. And as such it calculates feature importance of all features.



These are the feature importance based on Random Forest. **Price** has the highest importance followed by **ShelveLoc** and **CompPrice**.

1. Permutation method: An effective way to find feature importance is by using permutations. A model is trained multiple times on same data but by shuffling one feature's values multiple times. If that shuffle causes increase in error than feature is important otherwise that feature has minimal or no effect in predicting the target variable.

Here I tried permutation on KNN Regressor, showing the feature importance as-



Here as we can see **Price** again has highest importance, followed by **Population** and **CompPrice**.

1. One important observation can be made using these feature importance metrices, that features **Urban** and **US** have negligible importance, and making almost no impact on predictions, so can be dropped from data while training a model.
2. Linear Regression coefficients can also be taken as the measure of feature importance, the higher be the coefficient of a variable - the higher importance it carries.