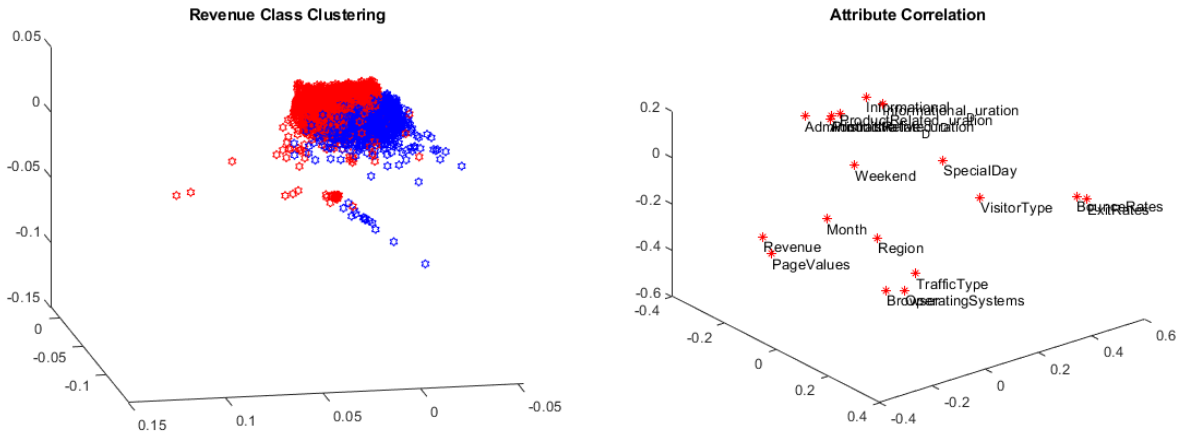
1 Clustering

Both k-means and Singular Value Decomposition (SVD) methods of clustering were used to determine more properties from the data, using only the important attributes discussed in ??.

1.1 Singular Value Decomposition

SVD clustering was attempted first to help visualize potential clusters and to further analyze the correlation between different attributes. From fig. 1.1a, there is an evident separation between data points in either of the Revenue's classes. So clustering using k-means should be feasible given the observed clumpiness, but may require multiple clusters due to how close the groups are. The Figure also shows promise for an accurate prediction model given the grouping of the two classes.



(a) SVD clustering of data points coloured based on Revenue class.
Blue = TRUE, Red = FALSE.

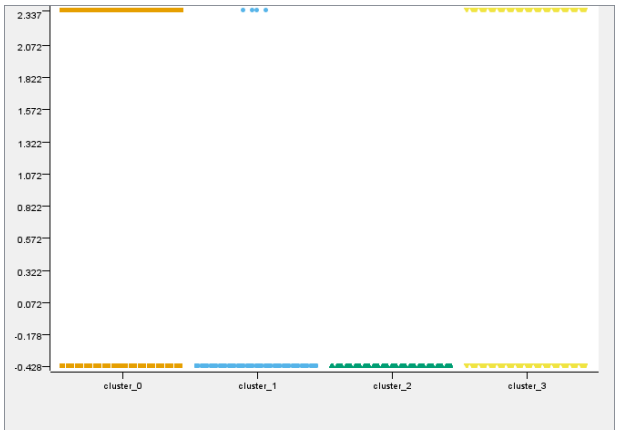
(b) SVD of the attributes.

Figure 1.1: SVD figures generated from a modified version of the `showsvd.m` MATLAB script.

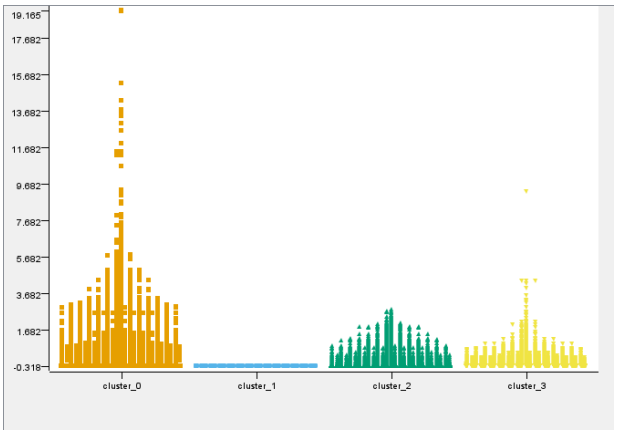
In addition, fig. 1.1b shows that the Revenue and Page Values attributes are highly correlated, which reinforces the correlation analysis conducted in chapter 1 and the attribute ranking done in ??. There is also some strong correlation between the important attributes located in different "correlation" clusters. The more notable of those clusters is the Bounce Rate and Exit Rate attributes being isolated and correlated similarly to the Revenue and Page Values attributes. This has some implications for the website's page structure, which will be further discussed in ??.

1.2 k-Means

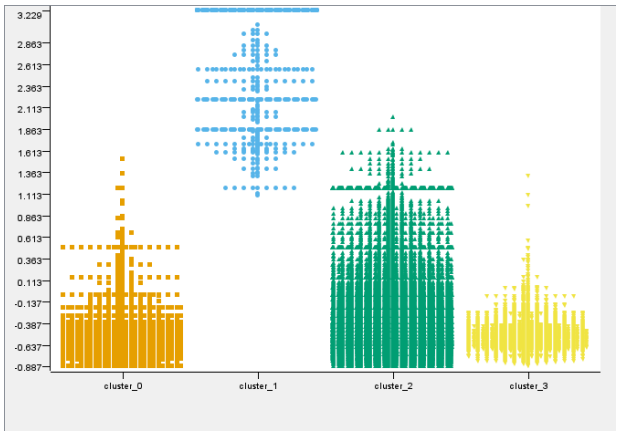
After trying various values of k in the range of $[2, 7]$, the number of clusters that provided the most useful insight was 4. From fig. 1.2, we can see that the clusters 0 and 3 reveal that revenue generating visitors do tend to have high Page values and low Exit Rates associated. The reverse trend can be observed in clusters 1 and 2, which reinforces the correlation analysis conducted in chapter 1.



(a) Clusters vs. Revenue



(b) Clusters vs. Page Values



(c) Clusters vs. Exit Rates

Figure 1.2: Scatter plots of the k-means clusters against significant attributes.

One may want to target clusters 1 and 2 to determine what would make them purchase something from Nozama's website, which will be expanded on in **??**.