DTSA 5510 Final Project

Big Mart Sales

Project Topic

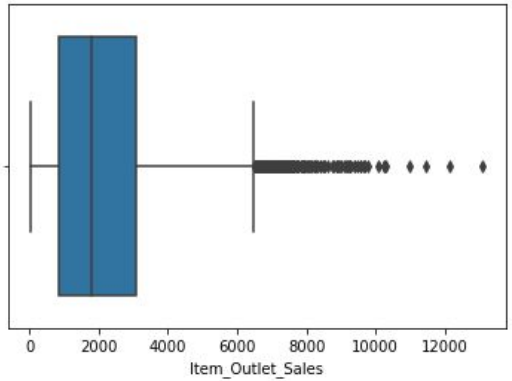
In this project I will be using the Big Mart Sales Data to solve the problem of customer segmentation utilizing unsupervised learning. The goal of this project is to be able to identify groups of customers that exhibit similar or distinct purchasing behaviors. This type of analysis could help Big Mart to create customized and targeted campaigns for specific groups and improve what products they offer to certain groups. To answer to problem of customer segmentation, I will be using unsupervised learning techniques like K-Means clustering and PCA. These will be valuable tools for BigMart to better understand their customers.

Data Cleaning/ EDA

Seems like Item_Weight and Outlet_Size have missing values.
Lets handle those missing values through imputation. Now we got rid of the missing values.

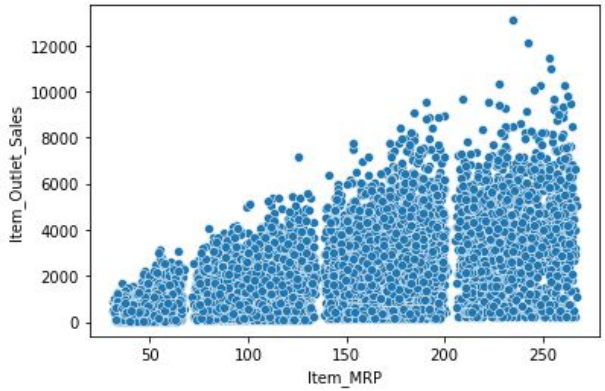
Now lets visualize for outliers.

```
# Check for outliers
sns.boxplot(x='Item_Outlet_Sales', data=train_data)
plt.show()
```



Looking at trends with a scatterplot of Item_MRP and Item_Outlet_Sales

```
sns.scatterplot(x='Item_MRP', y='Item_Outlet_Sales', data=train_data)
plt.show()
```



Analysis

Feature Selection

To select the relevant features that capture the key aspects of customer behavior I will need to analyze the data and identify the variables that are most relevant to the problem I am trying to solve. In this case, I want to identify groups of customers that exhibit similar/distinct purchasing behaviors. Therefore, lets select variables that are related to customer purchases.

Standardization/PCA

I'll need to standardize the selected features so that they have the same mean and variance. Creating this standardization step is important because K-Means clustering is sensitive to the scale of the data. I can use the StandardScaler function from the scikit-learn library to standardize the data. PCA process I'll be create the PCA class with the desired # of components. Then I will need to fit the model to the scaled data created before this step. After that I'll need to transform the data to the new PCA space.

Elbow method/Kmeans

To determine the optimal number of clusters for the data I'll be using the elbow method. The elbow method involves plotting the within-cluster sum of squares (WCSS) as a function of the number of clusters. After doing that then will be selecting the number of clusters at the "elbow" of the plot. we can apply the K-Means algorithm to the standardized data with 2 clusters

Results:

Cluster	PC1	PC2
0	-1.204525	-0.168660
1	2.973808	0.416398

The PCA analysis and K-means clustering have identified two distinct clusters of customers based on their purchasing behaviors. Cluster 0, which is characterized by negative values for PC1 and PC2, represents customers who make lower value purchases less often. While Cluster 1 is characterized by positive values for PC1 and PC2 and represents customers who make higher value purchases more often.

This information is useful for targeted marketing campaigns or tailored product recommendations. For example, businesses could use this information to offer promotions or discounts to customers in Cluster 0 to encourage them to make more frequent purchases or to focus on providing high-value products or services to customers in Cluster 1.

Conclusion

In conclusion, the Big Mart Sales data problem focused on the task of customer segmentation, which involves identifying groups of customers with similar/distinct purchasing behaviors. First started by selecting relevant features and then standardized them using the StandardScaler function. Next used PCA to reduce the dimensionality of the data and transform it to a new space. Then performed K-Means clustering to identify distinct clusters of customers. Determined the optimal number of clusters using the elbow method. Lastly, analyzed the resulting clusters using visualization and cluster analysis.

Customer segmentation utilizing techniques like K-Means clustering and PCA can be a valuable tool for businesses like BigMart to better understand their customers and tailor their marketing campaigns and product offerings to specific customer segments.