Data Science Assignment 3 – Matti Sevitt

1. **Perform EDA on the dataset to gain insights and understand the customer attributes. Include the following steps:**

   a. **Data cleaning: Handle missing values, outliers, and ensure data integrity.**

   There we no missing values, outliers or data integrity issues in the data set, thus no cleaning to be done.

   b. **Statistical summaries: Compute descriptive statistics (mean, median, standard deviation, etc.) for numerical variables and frequency distributions for categorical variables.**

   Descriptive statistics for numerical variables: (customer id is obviously irrelevant)

	Customer_Id	Age	Annual_Income_(k$)	Distance_to_Store_(km)	Online_Shopping_Score	Purchase_Diversity	Online_Engagement_Score	Social_Media_Influence	Customer_Lifetime_Value	Discount_Sensitivity
mean	150000.00000	44.02499	149.57029	15.47778	50.60012	10.48733	50.13974	49.92964	50.00768	49.98677
median	150000.00000	44.00000	149.00000	15.00000	51.00000	10.00000	50.00000	50.00000	50.00000	50.00000
std	28867.94647	15.30034	57.69277	8.06844	28.86488	5.76370	29.07487	29.18483	29.07922	29.17121

Frequencies for categorical variables:

```
Gender
Male                        33.50566494335057 %
Female                      33.32366676333237 %
Other/Prefer not to say     33.17066829331707 %
```

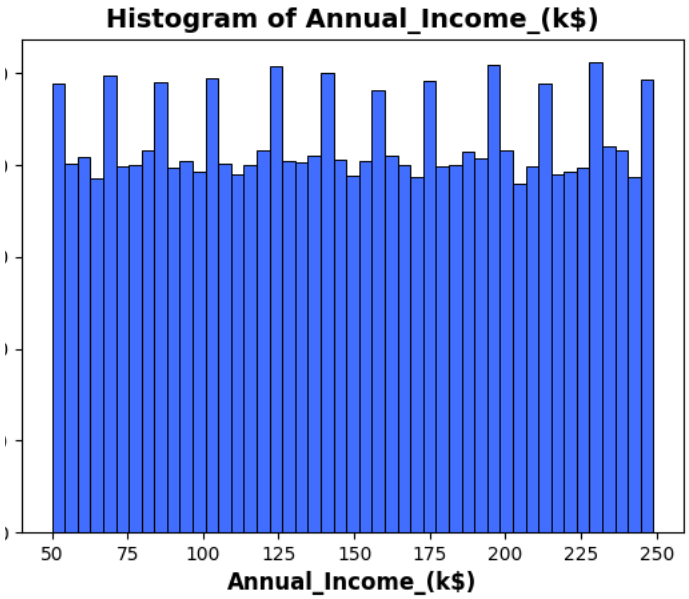
```
Education_Level
Postgraduate    25.185748142518577 %
High School     25.095749042509574 %
College         24.969750302496976 %
Graduate        24.748752512474876 %
```

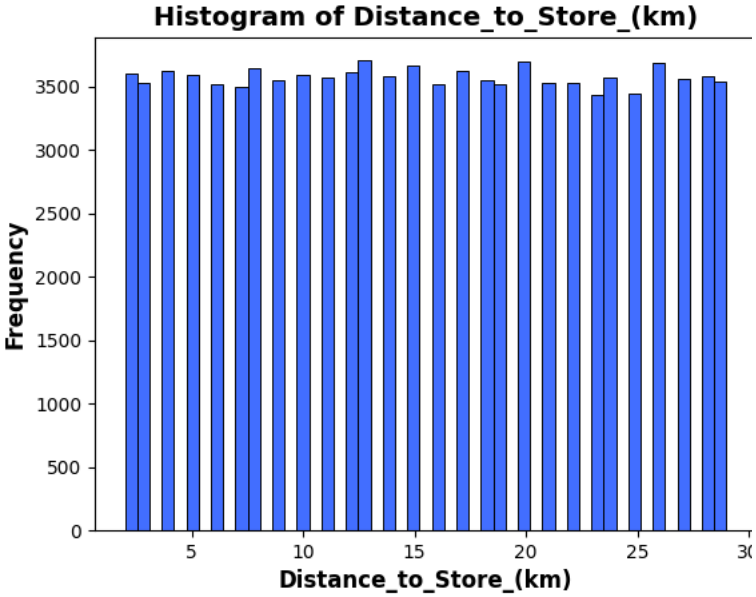
```
Marital_Status
Widowed     25.202747972520275 %
Single      25.045749542504574 %
Married     25.012749872501278 %
Divorced    24.738752612473874 %
```

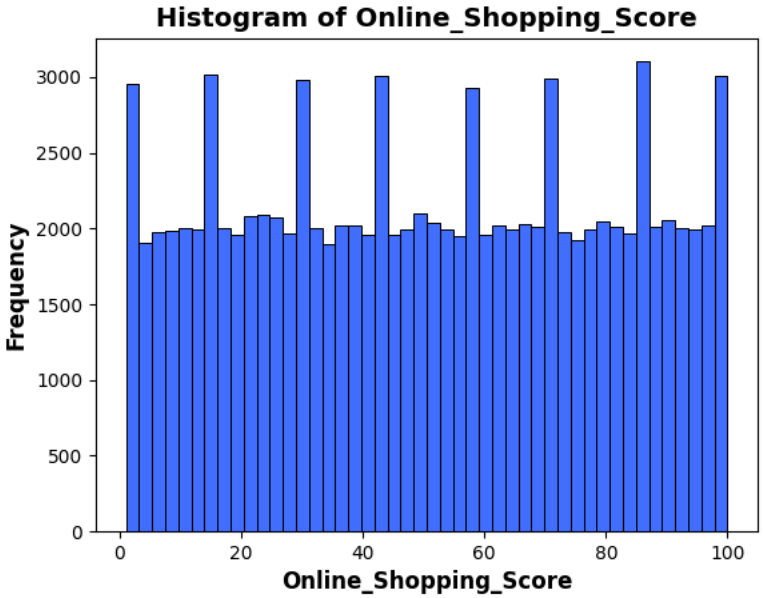
```
Number_of_Children
Two                 25.187748122518776 %
No children          25.00574994250058 %
Three or more       24.972750272497272 %
One                 24.833751662483376 %
```

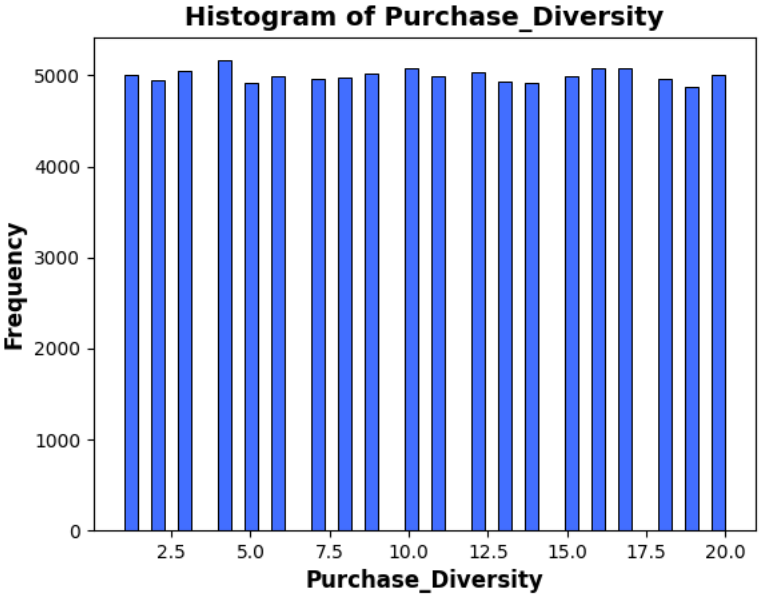
```
Minimalist            20.19579804201958 %
Brand Conscious       20.10079899201008 %
Impulsive Shopper     19.93080069199308 %
Bargain Hunter        19.92280077199228 %
Trendsetter          19.849801501984977 %
```

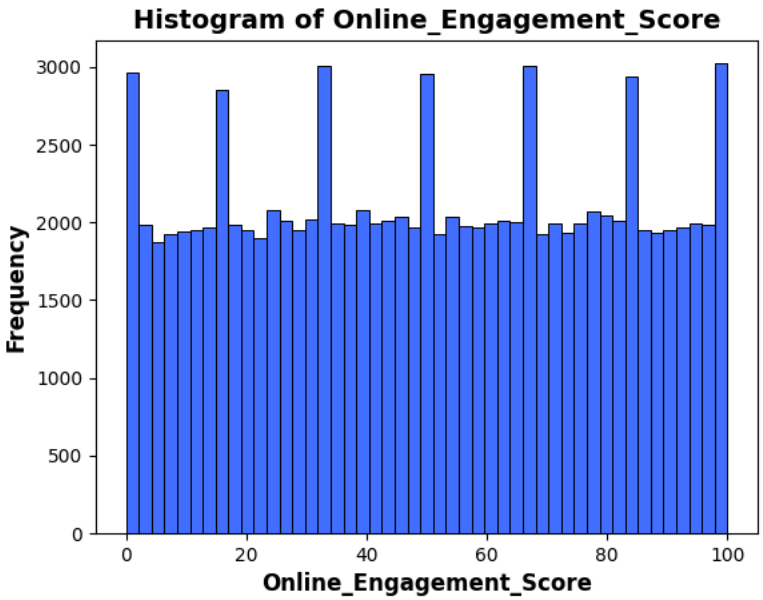
c. **Visualization: Create visualizations (e.g., histograms, box plots, scatter plots, etc.) to explore relationships and distributions among variables.**

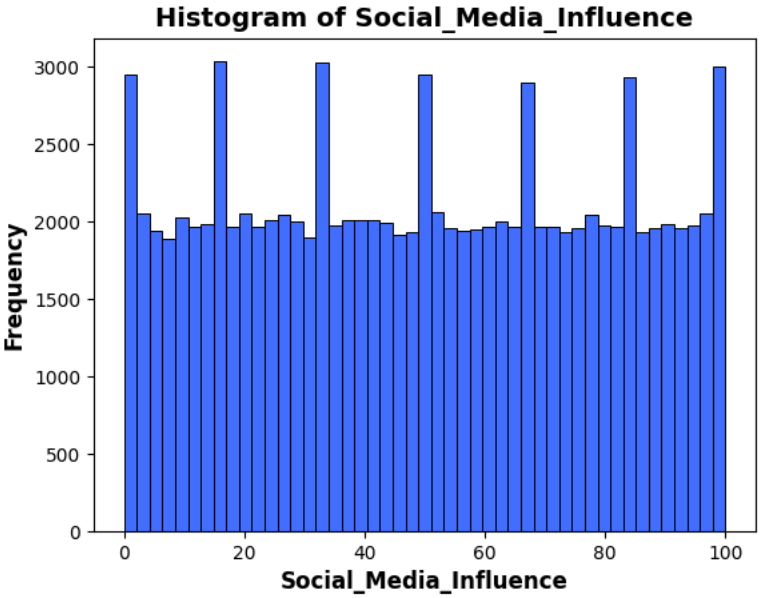


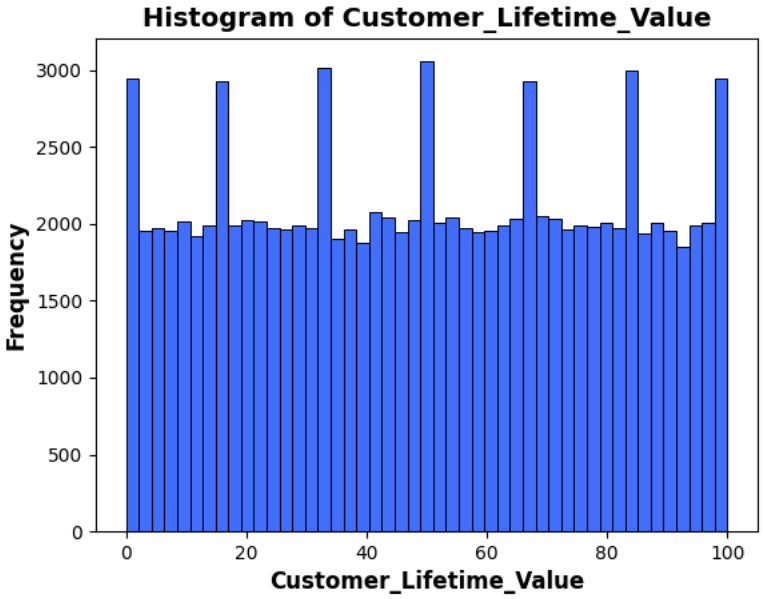


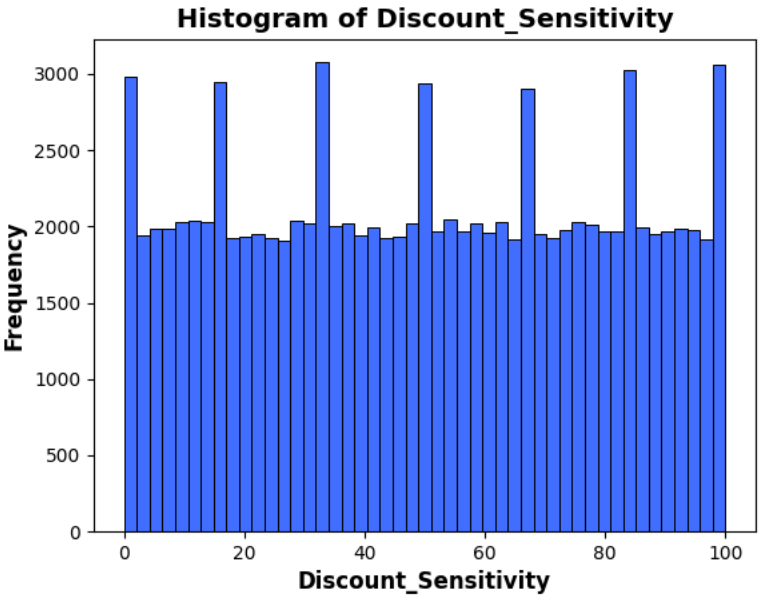


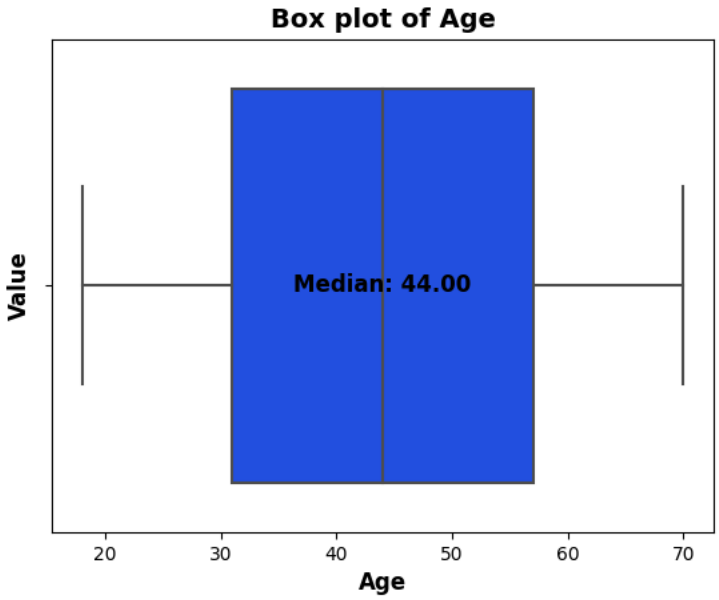

Histogram of Purchase_Diversity
Frequency
Purchase_Diversity
0
1000
2000
3000
4000
5000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

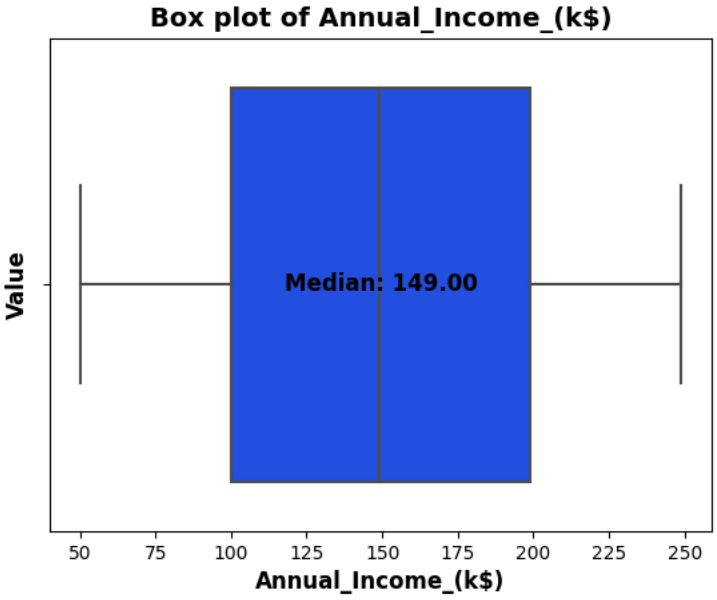

Histogram of Online_Engagement_Score
Frequency
Online_Engagement_Score
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100

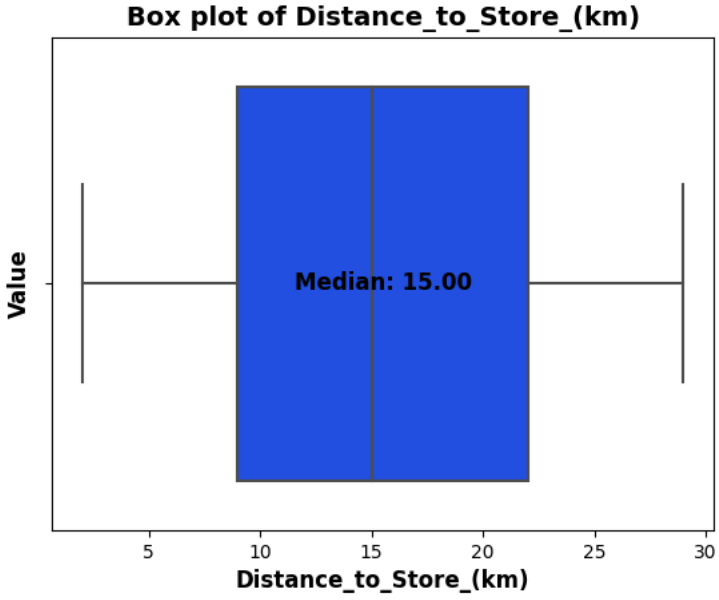

Histogram of Social_Media_Influence
Frequency
Social_Media_Influence
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100

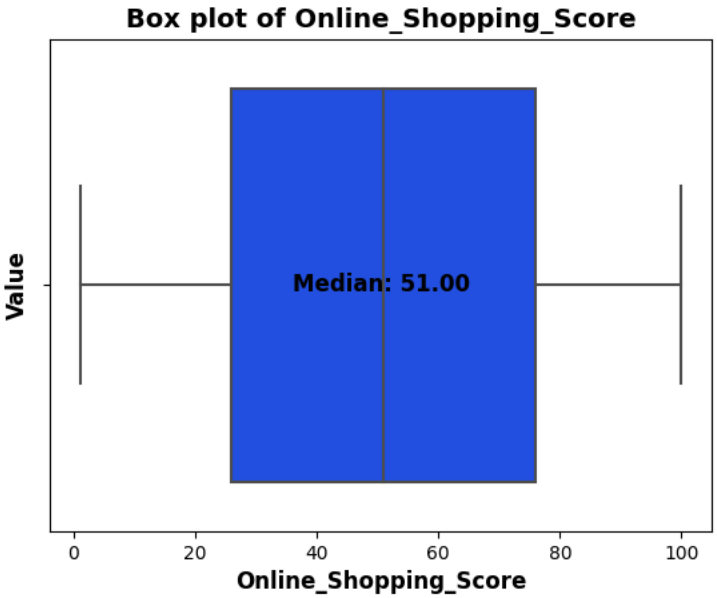

Histogram of Customer_Lifetime_Value
Frequency
Customer_Lifetime_Value
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100

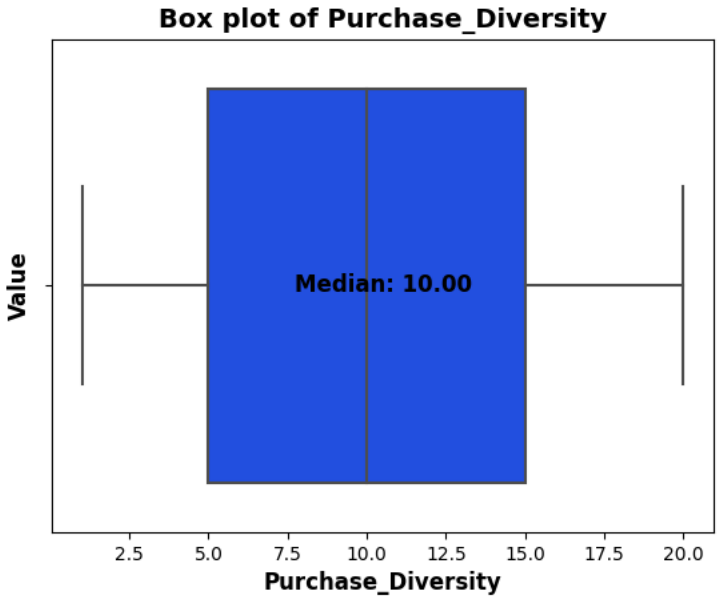

Histogram of Discount_Sensitivity
Frequency
Discount_Sensitivity
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100

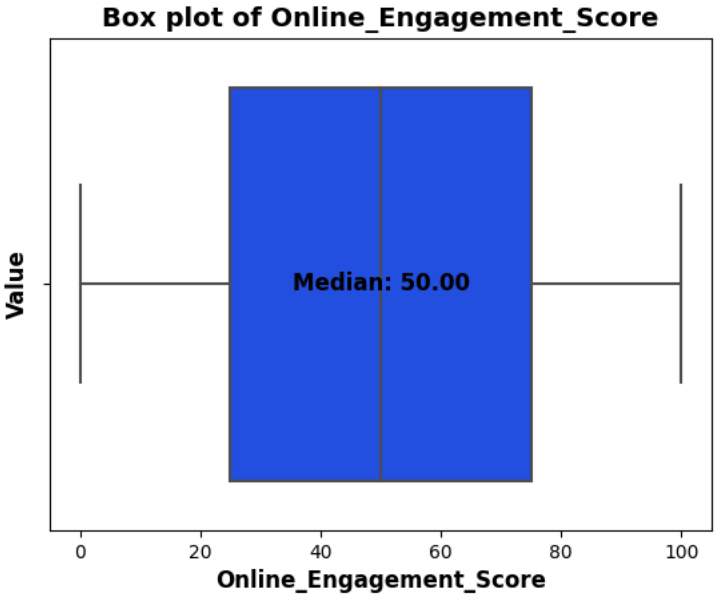

Box plot of Age
Value
Median: 44.00
Age
20
30
40
50
60
70

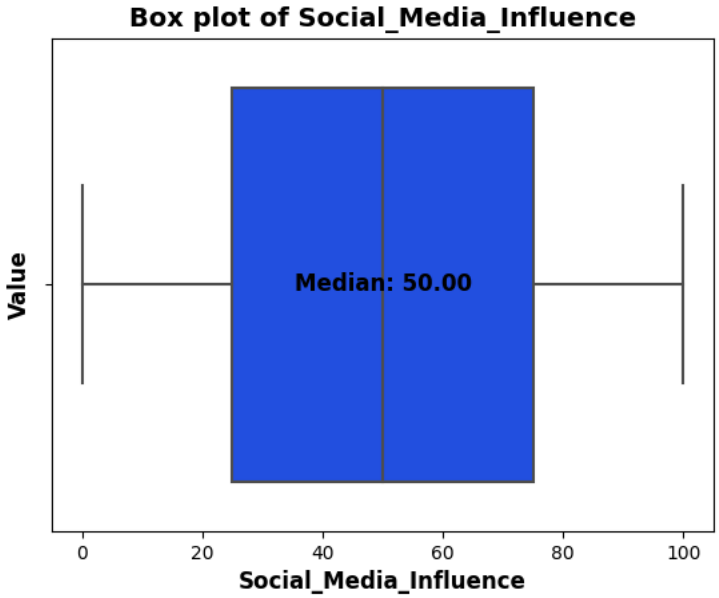

Box plot of Annual_Income_(k$)
Value
Median: 149.00
Annual_Income_(k$)
50
75
100
125
150
175
200
225
250

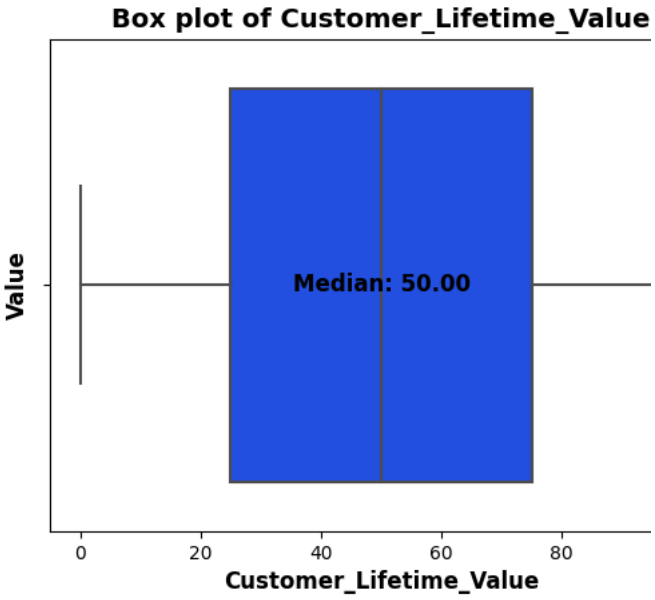

Box plot of Distance_to_Store_(km)
Value
Median: 15.00
Distance_to_Store_(km)
5
10
15
20
25
30

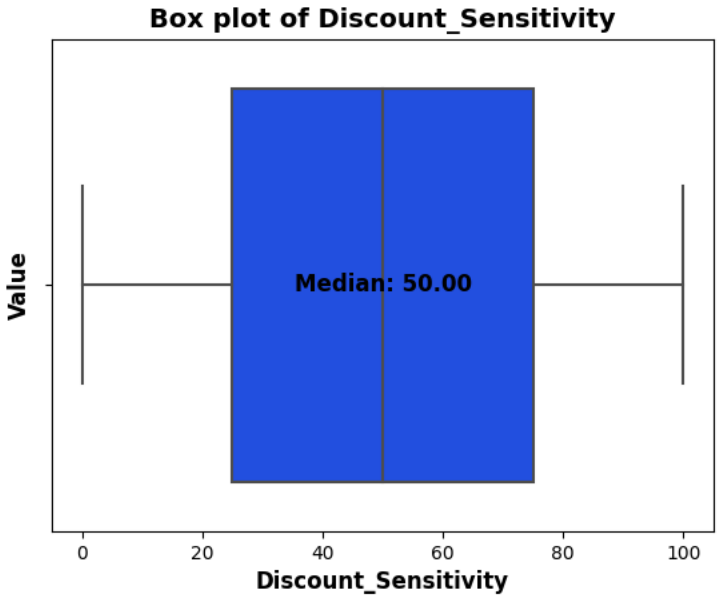

Box plot of Online_Shopping_Score
Median: 51.00
Value
0
20
40
60
80
100
Online_Shopping_Score


Box plot of Purchase_Diversity
Median: 10.00
Value
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Purchase_Diversity

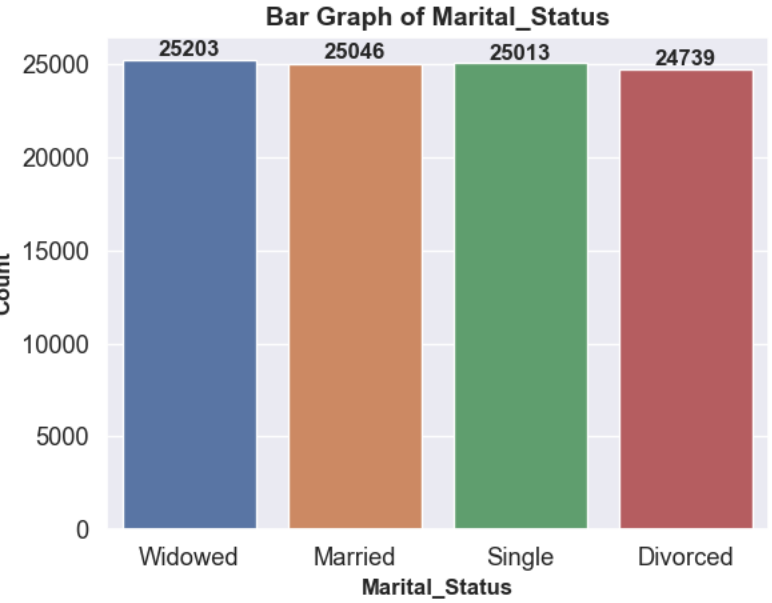

Box plot of Online_Engagement_Score
Median: 50.00
Value
0
20
40
60
80
100
Online_Engagement_Score

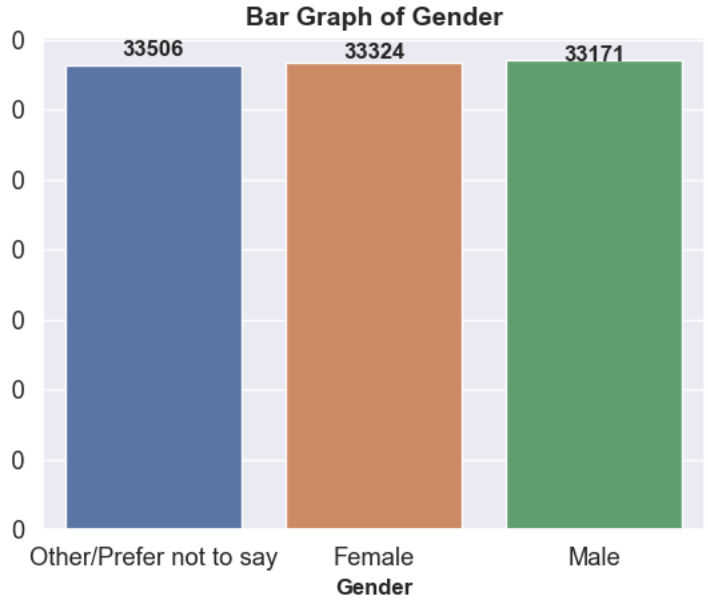

Box plot of Social_Media_Influence
Median: 50.00
Value
0
20
40
60
80
100
Social_Media_Influence

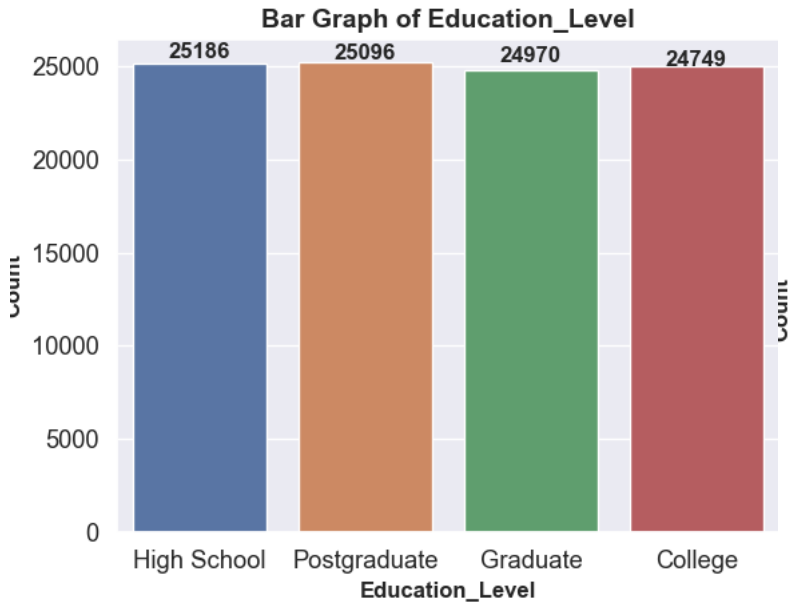


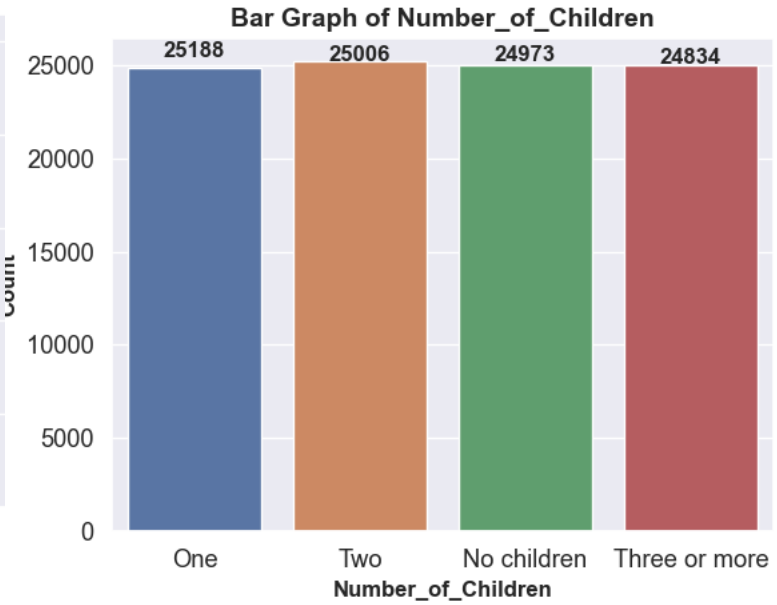


Correlation Matrix









As you can see in all the graphs the data is pretty much evenly distributed, there is no significant correlation between the different features and honestly no real conclusions from the visual data exploration. In some cases you can see slight differences, for example for a liitle bit more females than males in the data set, the median age of the samples is 44 while the age range is from around 20 to 70

2. **Apply clustering algorithms to segment the customers based on their attributes. Choose at least one clustering algorithm (e.g., K-means, hierarchical clustering) and perform the following steps:**

    a. **Data preprocessing: Standardize or normalize the numerical variables, and encode categorical variables.**

    I used the min max scaler and converted all the categorical columns to dummy variables.

    b. **Select the optimal number of clusters: Utilize appropriate techniques (e.g., elbow method, silhouette score) to determine the optimal number of clusters.**

    I have chosen to use the k-means algorithm to segment customers based on their attributes. I tried out different k values between 2 and 20 and concluded that 11 clusters has the highest score ((0.12312..) and therefore has the best clustering results.

    c. **Perform clustering: Apply the chosen clustering algorithm to the preprocessed dataset and obtain the clusters.**

    d. **Analyze and interpret the clusters: Analyze the characteristics and patterns of each cluster, and provide insights on customer segmentation based on the clustering results.**

    Most the clusters were similar in size (around 8000) while the second cluster was twice the size, The only significant difference that can be seen in this cluster is that the samples in this cluster have higher a marital and widowed status than all the other clusters, and none of them are single.

```
Cluster 0: 8356 samples
Cluster 1: 16628 samples
Cluster 2: 8390 samples
Cluster 3: 8505 samples
Cluster 4: 8340 samples
Cluster 5: 8235 samples
Cluster 6: 8315 samples
Cluster 7: 8308 samples
Cluster 8: 8194 samples
Cluster 9: 8313 samples
Cluster 10: 8417 samples
```

    This is my cluster analysis:

Cluster 0: The primary gender in this cluster is female with a high proportion having three or more children. This group is heavily influenced by social media, has a good online shopping score and shopping diversity. Their level of education is well-balanced. They are least likely to be bargain hunters and most likely to be brand conscious or impulsive shoppers.

Cluster 1: This cluster consists of male customers, and a majority of them have one child. Their online shopping score, purchase diversity, and online engagement score are close to the average. This cluster has a lower tendency to be bargain hunters or impulsive shoppers. They have a high sensitivity to discounts and a significant portion are high school graduates.

Cluster 2: This cluster is also male-dominated, with no children. These customers have slightly above-average annual incomes, online engagement scores, and customer lifetime values. They have a strong preference for minimalist shopping and a significant portion have a high school level education.

Cluster 3: This female-dominant cluster has no children. They have a tendency towards brand consciousness and a high proportion are graduate-level educated. This cluster has a high online shopping score and online engagement score.

Cluster 4: Customers in this cluster have selected "Other/Prefer not to say" for their gender and have two children. Their online shopping score, purchase diversity, online engagement score, and customer lifetime values are above average. This cluster has a fair balance in terms of their shopping persona.

Cluster 5: This cluster is filled with male customers who are widowed with no children. These customers have above-average annual income and online shopping scores. They have a strong preference for minimalist shopping and a significant portion are high school graduates.

Cluster 6: The sixth cluster consists of female customers with two children. These customers have higher than average online shopping scores and customer lifetime values. These customers tend to be brand conscious and impulsive shoppers.

Cluster 7: This cluster consists of female customers with two children. They have a slightly above-average online shopping score and online engagement score. They are likely to be minimalist shoppers and a significant portion are college graduates.

Cluster 8: This cluster is also female-dominated with three or more children. These customers have high online shopping scores, purchase diversity, online engagement scores, and customer lifetime values. They are more likely to be brand conscious and less likely to be impulsive shoppers. They are mostly high school graduates.

Cluster 9: This cluster consists of male customers who are married with no children. These customers have above-average online engagement scores. These customers tend to be bargain hunters and a significant portion are high school graduates.

Cluster 10: This final cluster is female-dominated with one child. They have high online shopping scores, purchase diversity, online engagement scores, and customer lifetime values. These customers tend to be bargain hunters and impulsive shoppers and are mostly college graduates.

3. Summary

The report offers a comprehensive analysis of JustBuy's customer segmentation based on an exploratory data analysis (EDA) and K-means clustering. The objective is to gain insights into customer behavior to enable JustBuy to make more informed decisions and tailor its marketing strategies.

The study employed a combination of statistical analysis, data visualization, and machine learning. It began with a thorough EDA to understand the dataset's composition and observe patterns. Next, a K-means clustering algorithm was applied to group customers into distinct segments based on their attributes.

Key Findings and Insights:

1. Exploratory Data Analysis: The dataset was evenly distributed with no missing values, outliers, or data integrity issues. Key insights include a slightly higher representation of females and a median customer age of 44 years. However, no significant correlations between different attributes were found.
2. Customer Segmentation: The K-means clustering identified 11 distinct customer segments, with a slight prevalence of females, varying numbers of children, different levels of online engagement, and differing shopping personas.

   Recommendations:

   Based on the customer segmentation, the following targeted strategies are recommended:

1. Marketing and Promotions: Tailor marketing messages and promotions according to the preferences of each cluster. For example, clusters that showed a tendency for brand consciousness may respond well to advertising focused on premium brands, while clusters that leaned towards bargain hunting would appreciate discount-centric promotions.
2. Customer Retention: Customer lifetime value varies across clusters, pointing to the need for specialized retention strategies. Higher value clusters may benefit from loyalty programs and premium services, while others may need incentives to increase engagement.
3. Product Recommendations: Personalize product recommendations based on the specific behaviors observed within each cluster. Clusters leaning towards minimalist shopping may appreciate a curated selection of essential items, for instance.
4. Sales Forecasting: Utilize the customer segmentation for more accurate sales forecasting. This will allow JustBuy to manage its inventory effectively, particularly for products favored by larger clusters.

   In conclusion the customer segmentation provides valuable insights into customer behavior, preferences, and value. By leveraging this segmentation, JustBuy can enhance its marketing strategies, improve customer retention, and streamline its operations for better overall performance.