Introduction

This project will use statistical analysis to provide strategic elements for a customer retention marketing campaign for a leading e-commerce platform. The dataset named *CA1_business_dataset* has 8000 variables and 25 features. These contain various types of sales information and company characteristics such as Transaction Amount, Product Rating, Product Category Preference and Customer Satisfaction Score, among others. By cross-referencing this data, we will try to bring out characteristics that are not readily apparent to the naked eye.

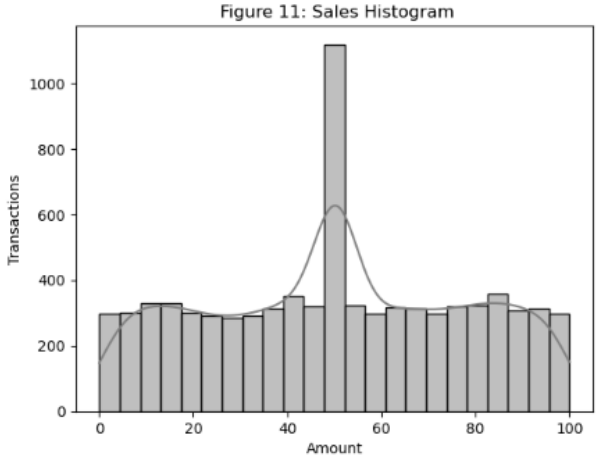
In an attempt to support itself with 5 background questions, the presentation of the data and its execution paths were carried out using the *Jupyter Notebook* platform which includes techniques such as descriptive statistics, probability variables and visualizations in order to be better understood.

Overview

Initially, they were divided into three large groups of relevant data: Pur**chase Frequency, Campaign Response Rate** and **Customer Feedback Score** and through them associated with features that most closely resemble these characteristics. It should be noted that, as the data is very similar, comparisons were tested with various features to try to find patterns and viable answers for the given context.

However, following the **Central Tendency Measures** that it was returned, we can describe some important info about it. In the first group of **Purchase Frequency**, it was identified that the average of each transaction was around 50.28 USD (as shown in images 11 and 12) and the **Standard Deviation** around 29% which suggests that the purchases are in a moderate scale within a reasonable range, and among the majority of consumer preferences, as shown in figure 7, **Gold** products stand out with 39% of interest and **Platinum** with 34%, showing a great aptitude for products in higher categories. Even though the maximum amount of transactions were around 100 and the minimum 0 are representing a large scale between both. This indicates that some customers have a high frequency of purchase like 100, but some of them don't even buy, which is representing zero. But the Standard Deviation strengthens the data that less than 30% is acceptable variation compared to all transactions.

Figure 11: Sales Histogram



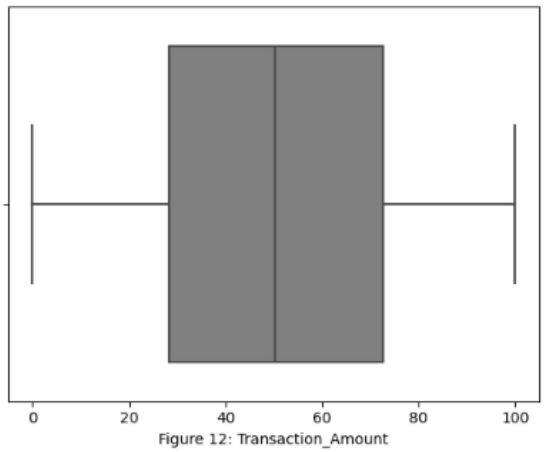


Figure 12: Transaction_Amount

Still with the same **population of 8,000**, the customers were divided into 4 large groups, all of which have similar samples of around 25% each, which shows a large division of the e-commerce company's customers. This data is strengthened when portrayed in terms of customer income. It can be seen that there is also a similar distribution of around 33% for **Low**, **Medium** and **High level** each, as shown in image 2. Therefore, there is a great decentralization of customers and their incomes, which indicates that the company does not have an approach or positioning to choose its focus customer.

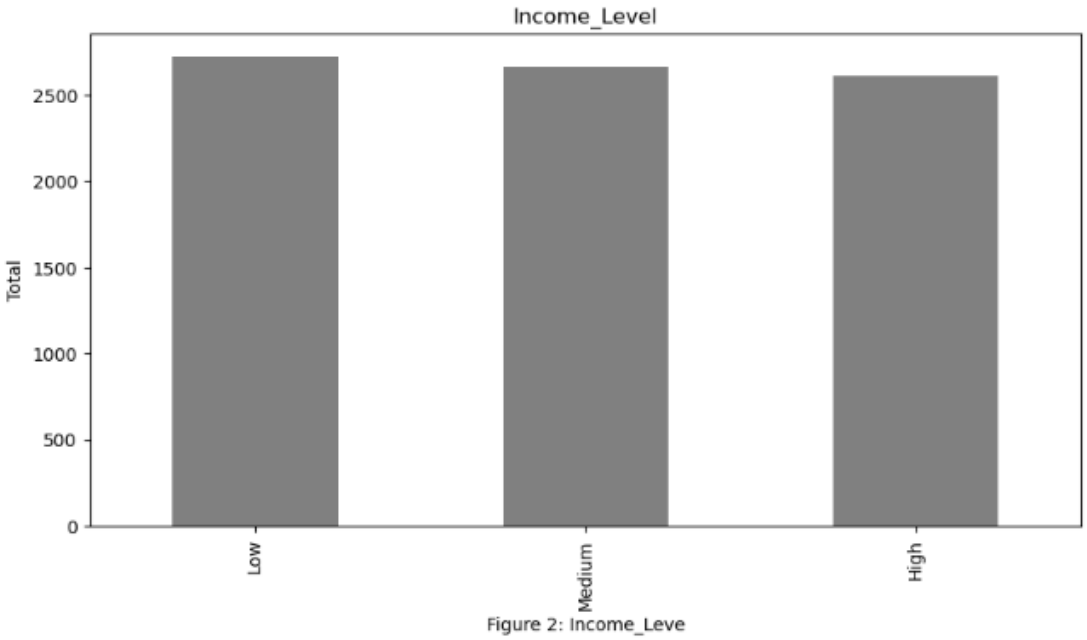


Figure 2: Income_Leve

Figure 7: Proportion by Product Category

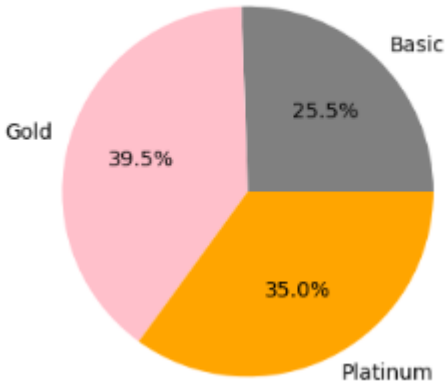
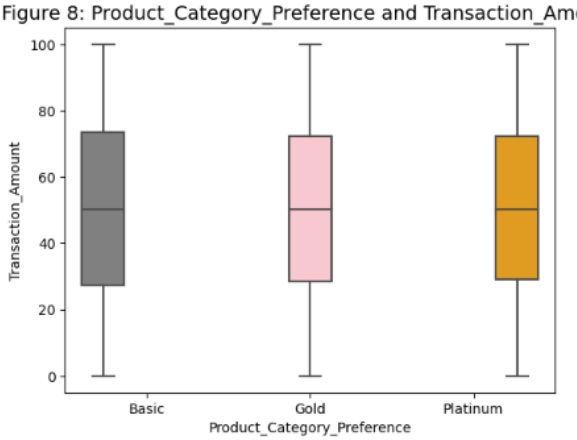


Figure 8: Product_Category_Preference and Transaction_Amount



When looking at the **High Value Purchase Count**, there are numerous transactions that are considered outliers, which means that there is a high dimensionality to the data and so, in order to return more reliable data, we tried to reduce this dimensionality. In addition, the greater concentration of high-value purchases returned data that on average is within the range of lower aggregate values,

comparing only the products that were sold at a high value, although there is a great deal of diversification of values and there is no product that stands out.

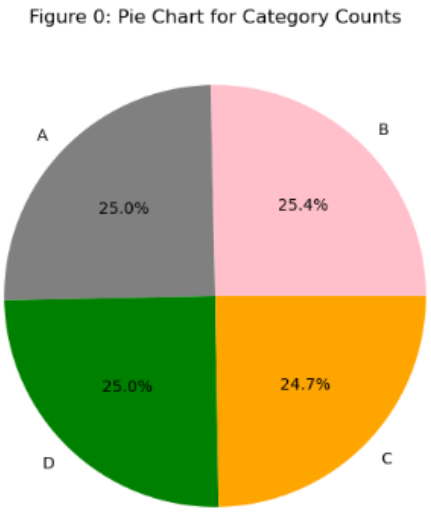


Figure 0: Pie Chart for Category Counts

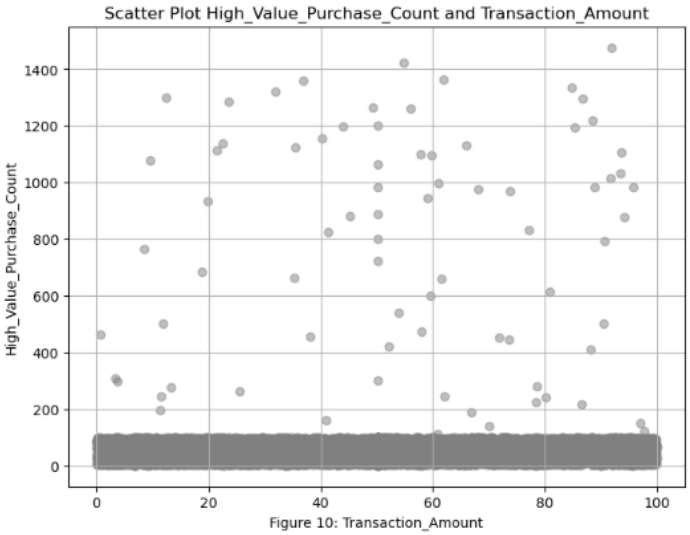


Figure 10: Transaction_Amount

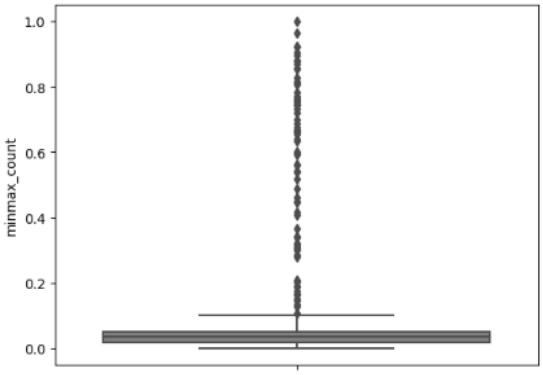


Figure 5: minmax

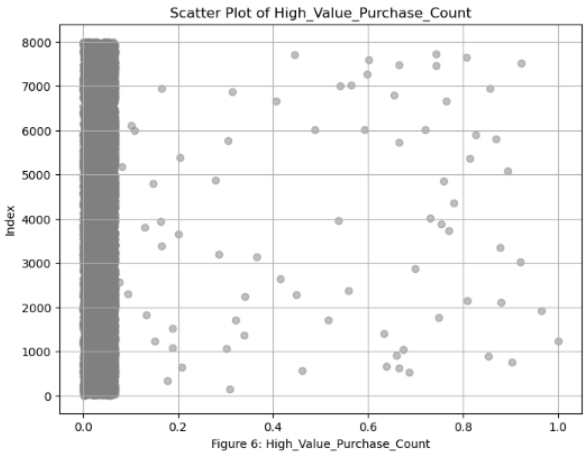


Figure 6: High_Value_Purchase_Count

Moving on to **Campaign Response Rate**, **Marketing Email Click Rate** data was measured, which shows that Marketing emails have an average click-through rate of 49%, although there are people who click frequently and others who don't, which seems natural. These figures are similar when compared to the **Social Media Engagement Score**, so the approach can be more effective in both mediums for reaching and engaging customers.

It was also calculated that 49% of the products had an upsell as shown in image 15, i.e. when a product was purchased, a related product was recommended. This figure shows a good return, indicating that there could be a greater approach with correlated items that could have a greater impact resulting in increased sales.

The number of customers interested in new products was around 50%, representing 4049 of the total population.

Figure 16: Percentage who has Interest in New Product

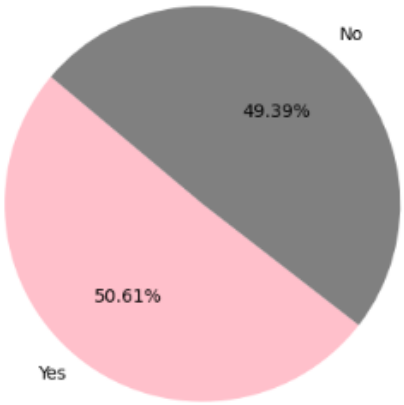


Figure 15: Product_Upsell_Success

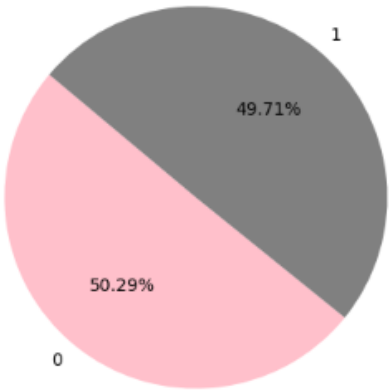


Figure 13: Product_Upsell_Success and Social_Media_Engagement_Score

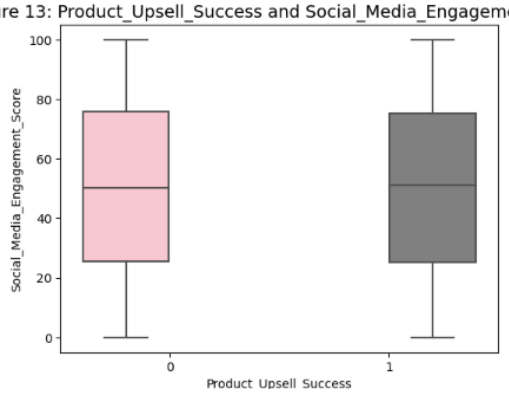
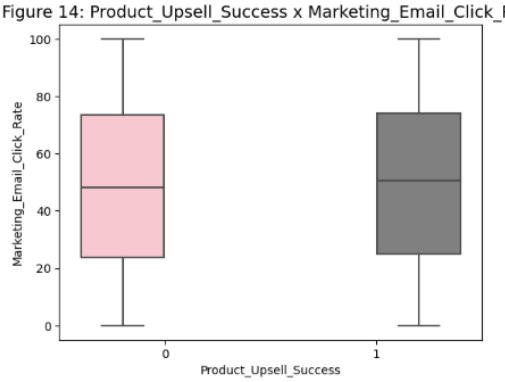


Figure 14: Product_Upsell_Success x Marketing_Email_Click_Rate



Now, highlighting the **Customer Feedback Score**, initially it had missing values in the **Product Rating** f eature, so the average (50.01) was applied to replace the zeros and be able to fit them on a scale of 1 to 5. The data returned that the average rating of the products given was within the range 3, which r epresents around **60** to **80** ratings, a good average which demonstrates good acceptance and satisfac tion with the products; and with a **Standard Deviation** of around **27%** which also means a good samp le with little variation between the averages, strengthening the data as shown in images 17 and 18.

In relation to the products correlated with the amount of transactions, they were also classified simil arly, but the products with classifications between 40 and 60 stand out, as they had lower Deviation v alues, i.e. less variation. Between the first and the third quartile the customers are in a range of 40 an d 60, yet there are few outliers. This shows that there is still room to improve the service, more direc tly focusing on products that have a price close to this range, as these are where the products have h ad the most sales and the most satisfaction.

The product reflects the final quality and satisfaction with the company, i.e. after buying a product, t he customer has good satisfaction, although it is still quite diverse between other products.

Figure 17: Product_Rating and Customer_Satisfaction_Score

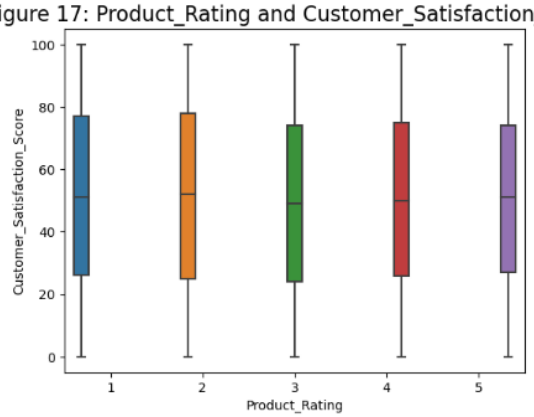
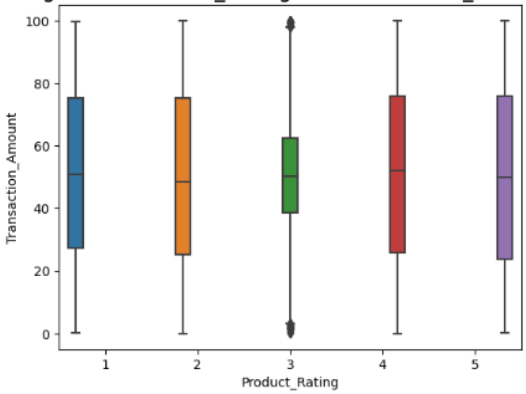


Figure 18: Product_Rating and Transaction_Amount



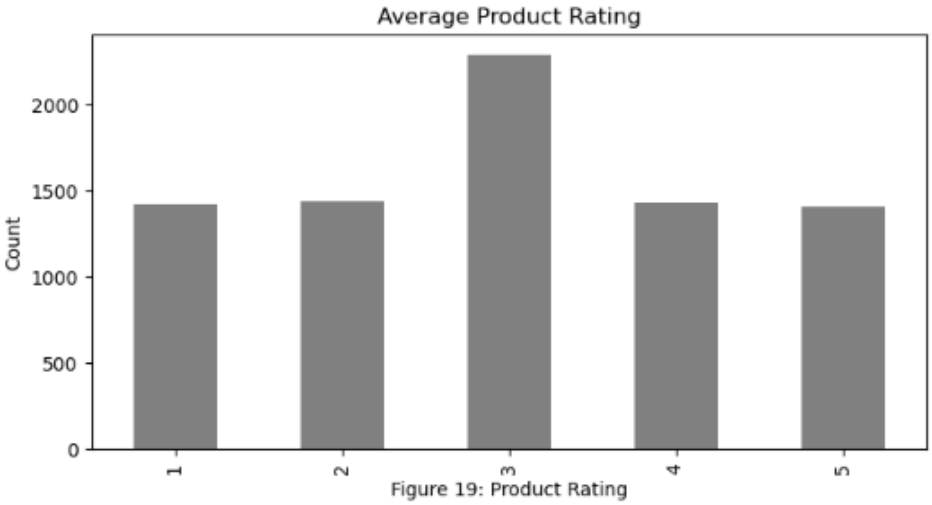
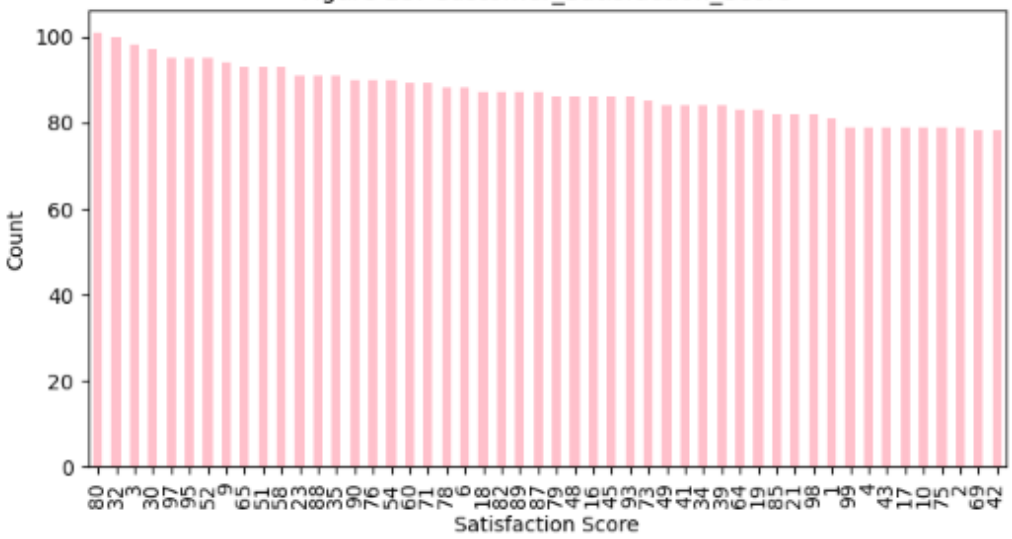


Figure 19: Product Rating

To finalize the simple descriptive approach before cross-referencing the data, we highlight the **Customer Satisfaction Score** without an exact definition, because although the Standard Deviation was around 28.98% - which brings little variability to the scores - it moves away from the average of 50.21 and with the first and third quartile the customers satisfaction in a range of 25% and 75%, which demonstrates a balance between the data that can be seen in the figure below.

Figure 23: Customer_Satisfaction_Score



General Initial Highlights

In summary, therefore, Gold products with 39% and Platinum products with 34% are the products most preferred by consumers with average transactions of around 50 USD, which shows that there is room to increase consumption of mainly Gold and Platinum products. An important attribute for the company's marketing team to take into it is that there is no definition of the segment of focus customers as A, B, C or D, and as well as their Low, Medium and High incomes, they all have great similarities.

It is also worth noting that there is no one product that stands out in terms of sales, but the 49% upsell indicates that there is a high percentage of people who buy products associated with the initial product that the customer is looking for. Therefore, the right approach of associating similar products with the purchase can lead to greater sales. And when associated with customers who are interested in new products, which stood at 50%, it shows a propensity for new purchases, so an approach with Gold and Premium products associated with similar products and a strategy that focuses on customers who are interested in new products could be viable for achieving greater consumer reach. Linked to this context, products that had a rating around scale 3 had a higher concentration of price per product around 40 and 60, while the others had similarity between 30 and 75. In terms of communication, both email marketing and social media can be considered, as both have similarity when returning a score from customers.

Highlights for probabilities

When crossing the data through Binomial distribution to obtain the probability of successful engagement with the marketing campaign, the Customer Segment Group variable was chosen and crossed with other categories such as Product preference category, Subscription status and Interest in new products.

In the first distributive crossover, we tried to select segment B to highlight 5% of the total 1998 sample, and also cross-referenced it with the preferred product Gold, which was one of the preferred products, as mentioned above. By choosing 50 as 6.5% of the sample, the result returned a 98% probability, when testing 40 as a random variable the result dropped to 61.9%, thus demonstrating that there is a 98% probability that 50 people or less of the sample size have preference for product B.

The second analysis highlights segment A customers with subscription status, where 0 represents unsubscribed and 1 subscribed. The aim is therefore to cross-check the data of people who are interested in receiving product information by associating them with segment A. For the calculation, the percentage of the sample size was between 50 and 70, representing between 5% and 7% of the sample. Although we tested between 50 and 60 and the result was close to 55.35%, the first analysis returned 57.53%, which can still be considered low, so there is a 57% probability that between 50 and 70 people are subscribed and belong to segment A. It should also be noted that the sample size was chosen from 5% of the total sample of A 1998, resulting in 100.

The third analysis sought to correlate customer segment B with interest in new products. So 'Yes' represents those who are interested in product B with a probability of between 50 and 80 people,

which represents around 5% and 7.7% each of the sample of those who are subscribed. The result returned 78.61%. Even though it was tested between 50 and 60, the result was lower at 71.68%. Therefore 78.61% of the first analysis shows a high probability that between 50 and 80 subscribers are from segment A.

As the end of this CA and as requested, the Customer Satisfaction Score variable was chosen, as it is associated with customer satisfaction after a purchase and can be considered a marker for whether customers are likely to buy products from the company again. Therefore, in order to find a Normal Distribution, $\mu = 50.21$ and $\sigma = 28.98$ were established for an initial probability of less than 40, which resulted in 36.22%; for greater than 40, it resulted in 63.77, and also between 50 and 80, which resulted in 35.08%, with the latter showing the lowest probability and greater than 40, the best. Therefore, an adequate approximation to the normal distribution was made by considering that 40, as shown in image 24, is very close to the values observed in the data. This suggests that the normal distribution is a good representation of the data and indicates a distribution that behaves as expected.

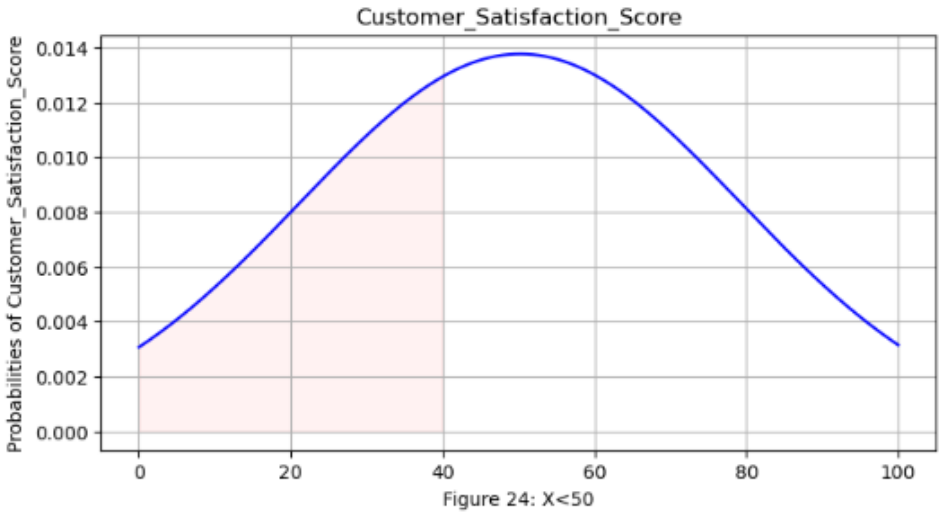


Figure 24: X<50

Finally, it can be seen that there is a difference between the Standard Deviation and the mean when comparing the Customer Satisfaction Score and High-Value Customers data, as the latter variable has numerous outliers which increases the Standard Deviation (58% approximately). The average was also below half of the values (38% approximately), so they do not have similar characteristics.