adbelghafar.hosni2019@feps.edu.eg

https://www.linkedin.com/in/abdelghafare-hosny-b64807240/

https://github.com/Abdelgafare/

Synthetic Data Simulation For Taager

Prepared by Abdelghafar Hosny Muhammed

Table of Contents:

Abstract

This research paper presents an in-depth analysis of Taager, a platform revolutionizing social e-commerce in the MENA region. Taager empowers entrepreneurs by providing a platform to start and grow their online businesses, fostering economic growth and innovation.

The study will leverage simulated data to gain insights into various aspects of the platform. The data includes variables such as Customer ID, Cost of Item, Day of Purchase, Day of Receiving, Day of Referral, Type of Shipping Company, Complaint Type, Product Category, Product, Shipping Duration, Product Review, Payment Method, Governorate, Quantity, Customer Profit, Total Profit BY Cust, Customer Satisfaction.

The analysis will involve the application of various data analysis techniques and machine learning models. Specifically, the study will conduct cluster analysis to identify patterns and segments within the data. A recommendation system will be developed to personalize the user experience on the platform. The purchasing behavior of customers will be studied to understand the factors influencing their decisions.

Furthermore, classification models such as Random Forest Classifier (RFC) and Logistic Regression will be applied to predict outcomes based on the data. The Python programming language, along with libraries like pandas, NumPy, matplotlib, seaborn, and sklearn, will be used for the analysis.

The findings from this study are expected to provide valuable insights that could help Taager further empower entrepreneurs and optimize its platform. The results could also contribute to the broader understanding of social e-commerce dynamics in the MENA region.

First: Simulation Process

The simulation process for the research paper on Taager was conducted using Python. The process involved generating a dataset with 1000 rows, each representing a unique customer transaction. The variables in the dataset were chosen to reflect key aspects of customer transactions and behaviors on the Taager platform.

The variables include:

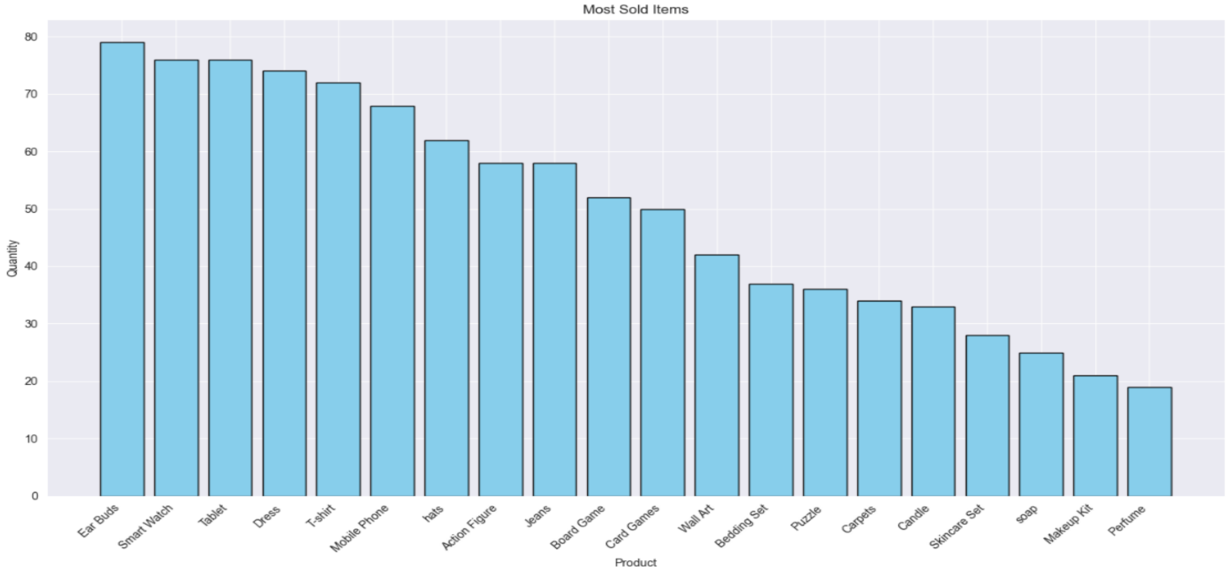
1. Customer ID: Randomly generated customer IDs.
2. Day of Purchase, Day of Receiving, Day of Return: Randomly generated dates representing the day of purchase, receiving, and referral.
3. Type of Shipping Company: Randomly assigned 4 types of shipping companies.
4. Complaint Type: Randomly generated 4 complaint types with different probabilities (Product Quality, Website issue, Shipping).
5. Product Category, Product: Randomly assigned product categories I choosed 5 Categories and 4 products based on the categories.
6. Shipping Duration: Calculated as the difference between the day of receiving and the day of purchase.
7. Product Review: Randomly generated product reviews (1-5).
8. Payment Method: I assigned 4 payment methods with different probabilities.
9. Governorate: Assigned 16 governorates in Egypt with different probabilities.
10. Quantity: Randomly generated quantities for each product.
11. Cost of Item: Updated based on the assigned product.
12. Customer Profit: Calculated based on the revenue the customer will get based on the cost of on product.
13. Total profit by customer: The Quantity X Customer Profit
14. Customer Satisfaction: Calculated based on the complaint type, product review, and shipping duration.

The probabilities for the random assignments were chosen to reflect realistic distributions. For example, the Payment Method variable was assigned with a higher probability for ‘Valu’ and ‘Cib bank’ to reflect their popularity.

The cost of each product and the profit for the customer were defined based on realistic estimations. The customer satisfaction score was calculated based on the complaint type, product review, and shipping duration, with higher scores indicating higher satisfaction.

Second: EDA (Exploratory Data Analysis)

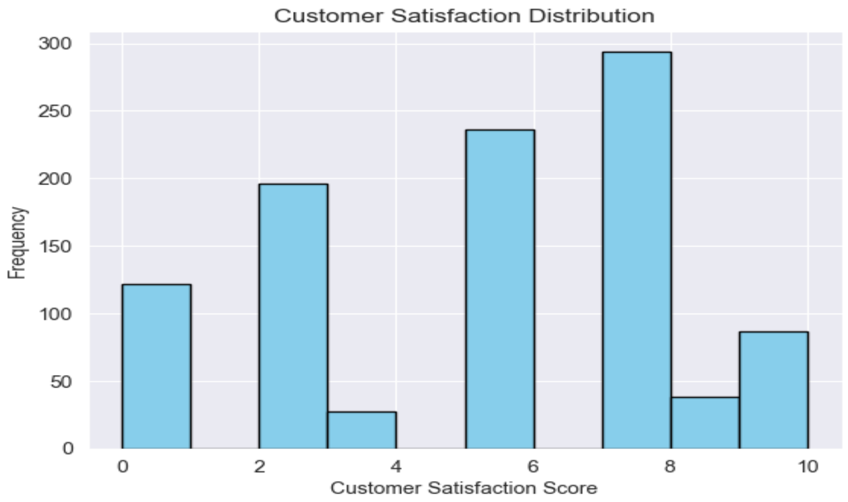
1. **Most Sold Products**: I’ve identified the most sold products in each category. This information can be useful to understand customer preferences and can guide inventory management. For example, the most consumed product the ear buds that I cannot dispense from it the inventory and try to enhance this product by reliability tests and model to prevent the returns and buying similar products



2. **Customer Satisfaction:** I’ve analyzed the distribution of customer satisfaction scores. This can help identify if there are any issues with customer satisfaction that need to be addressed.

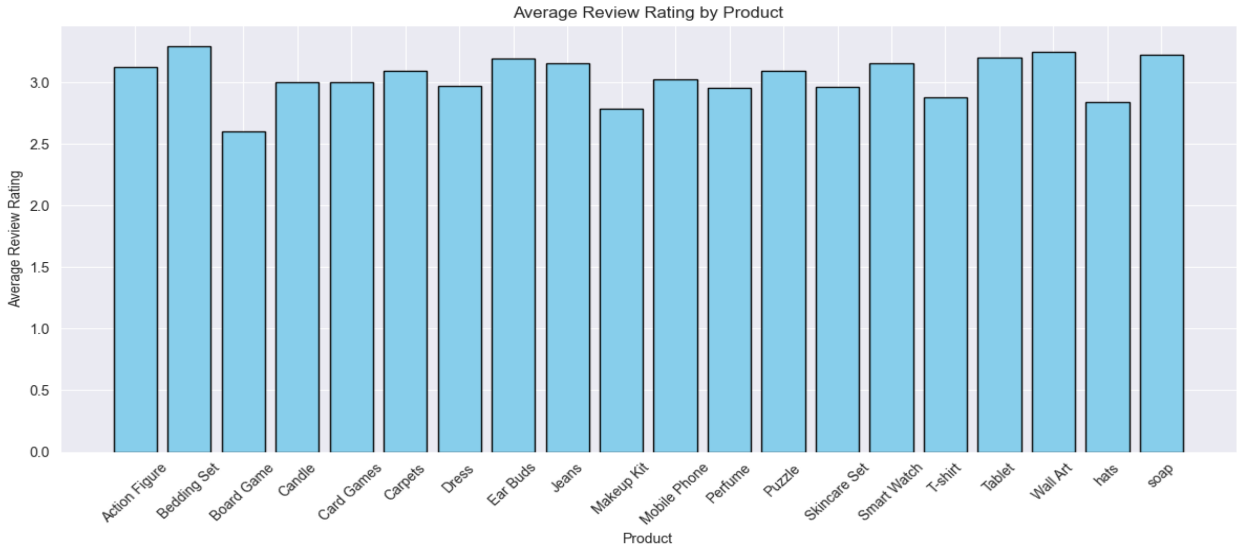
   I need to clarify that is overall customer satisfaction from the whole process and the satisfaction are based the complaint type, product review, and shipping duration.

   So, as we see here nearly like 300 people, we can call it satisfied from the whole process but there something didn’t make our score here for example equal to 10 so let’s break down this variable and try to study this behavior more accurately



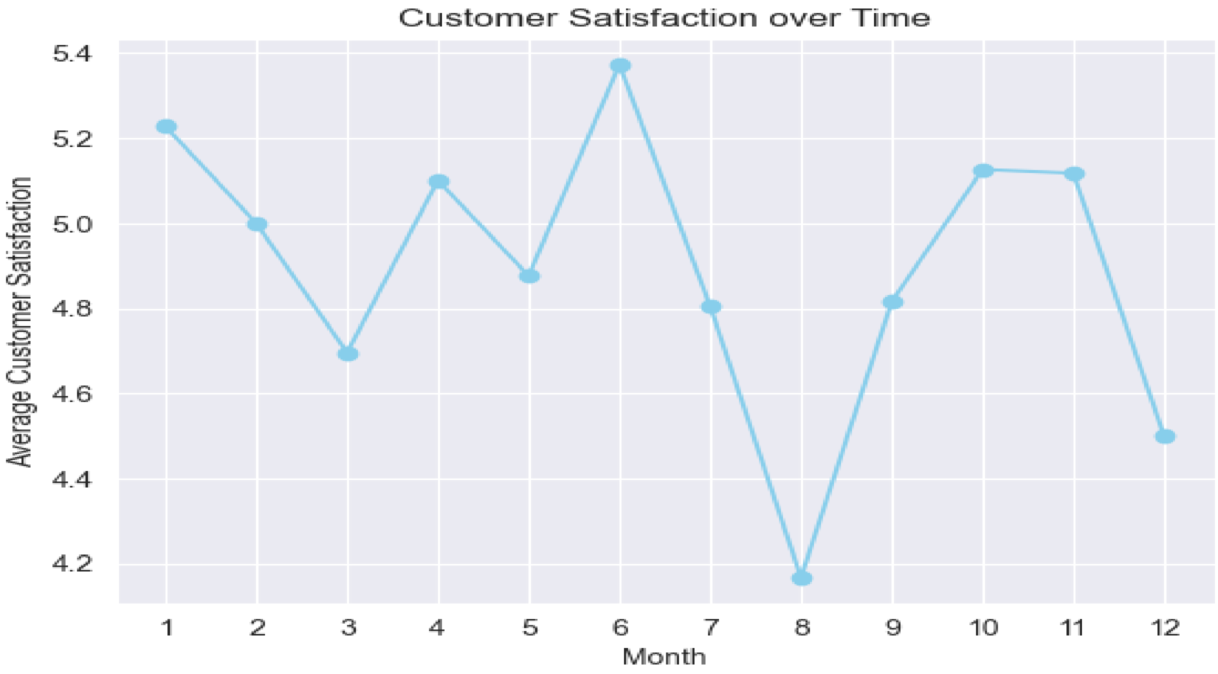
From the figure below we see that average reviews of all products varies between (2.5-3)

Which seem that product review looks moderate



From the figure below we see that customer satisfaction level have seasonality every month that mean that something wrong happens every end of the month and also have a linear trend across the year and the trend take place at point 5 from 10 which seem that customer stratification level is moderate and have a falling point at month 8

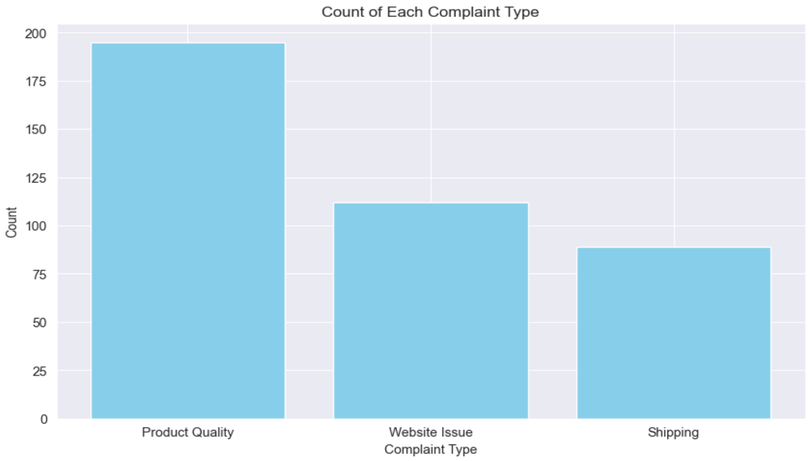
If we the trend is decreasing crucially from month 6 to 8 which indicates that there a problem was in this period and also the problem was fixed and get back to the normal at 5 which seems not good by the way we should get into to the factors that affect customer satisfaction level and try to fix it



3. **Complaint Types**: I’ve counted the number of each type of complaint. This can help identify common issues that customers are facing.

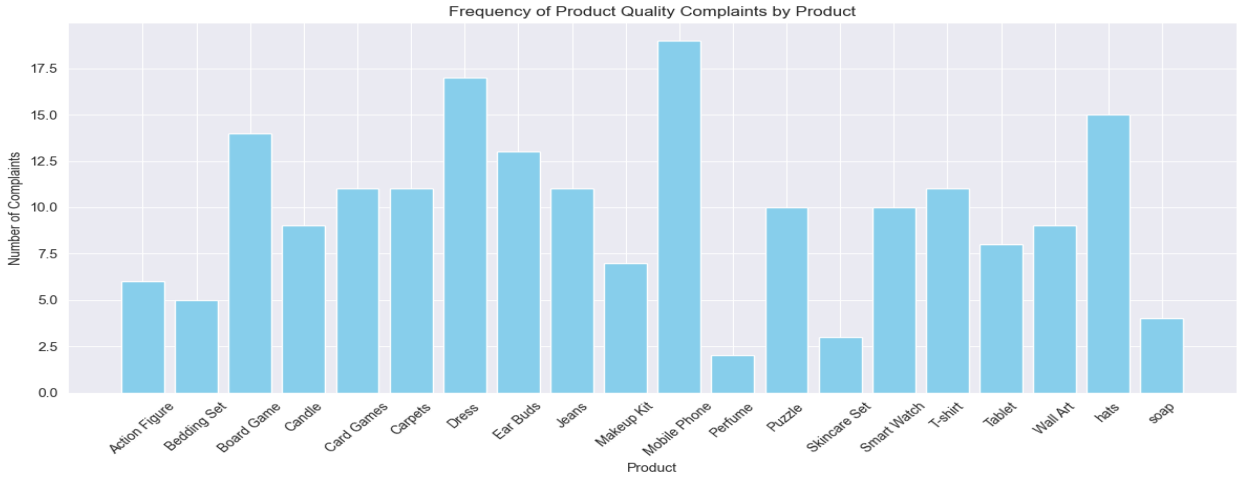
We see that the most complaints occur due to the Product Quality then the website issue so we are addressing issue right here and it’s the product quality so let’s get deeper and know which one of the products gets the most complaints about the quality, the website issue and the shipping seems equal too and we know as the company that maybe the website because the maintains

So, we will break down the both product quality and shipping to try improve those both weakness

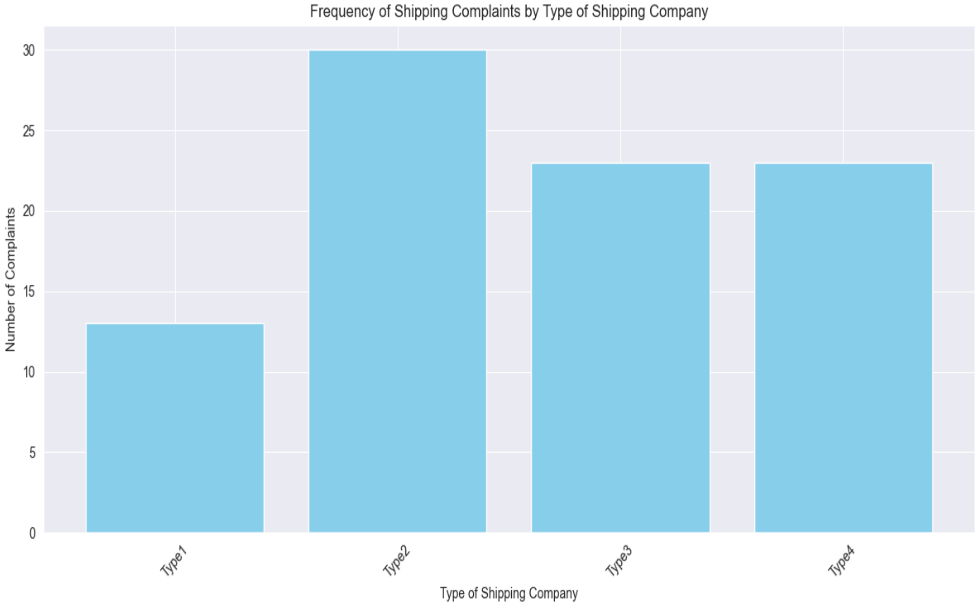


The figure below say that the mobile phone and hats and dress and board are the most products that got complaint about its quality

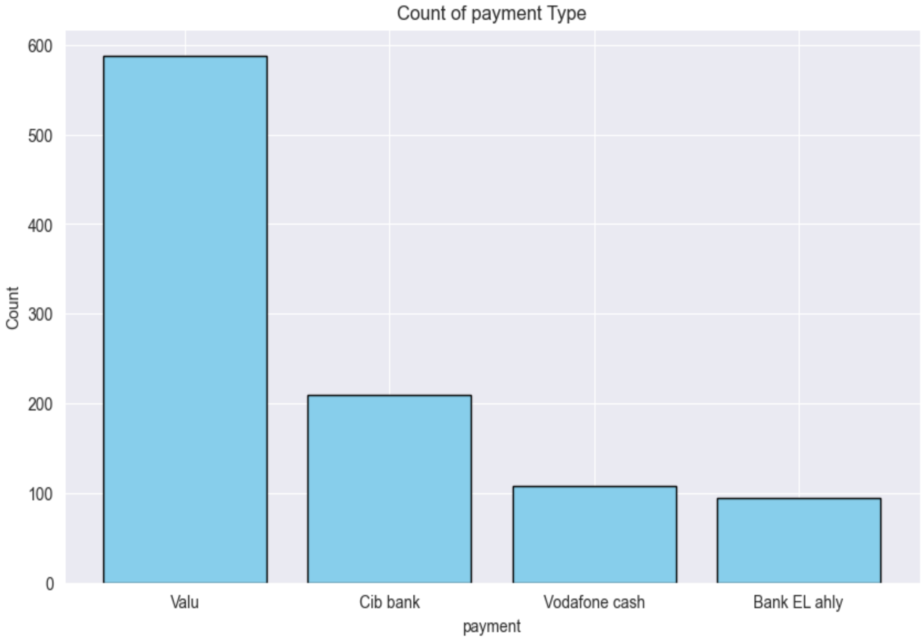
So, I suggest we handle this problem by dealing with another product that have higher quality



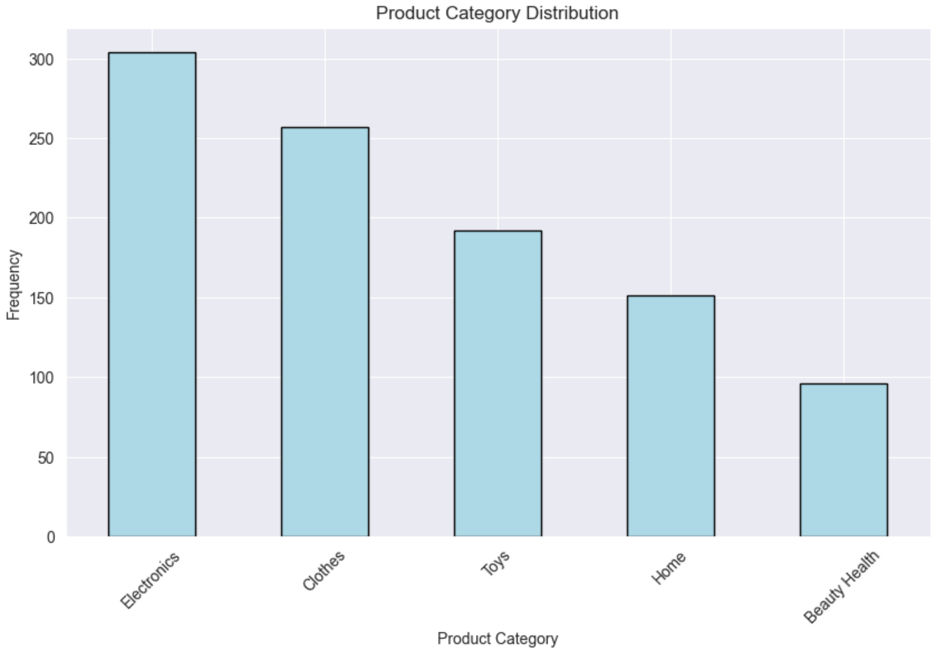
The figure below shows the relation between shipping complaints and the type of shipping company and we see that the type 2 shipping are the most one to claim the complaints about the shipping



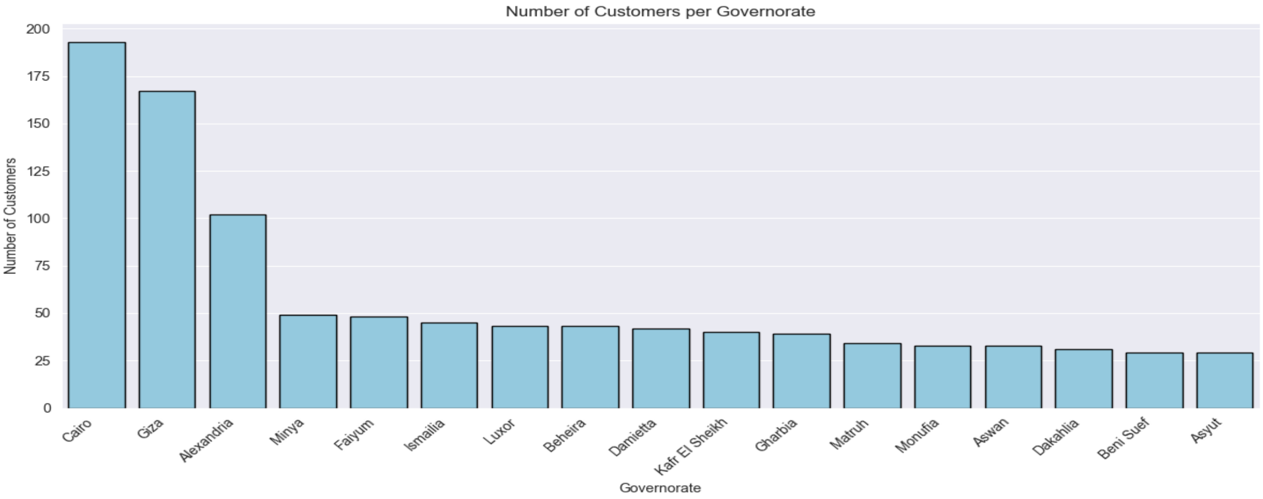
4. **Payment Methods**: I've determined the count of each payment type. This can provide and it look obvious that Valu is dominating in the payment maybe because it's easier way to pay or they can provide you better options to pay by installments



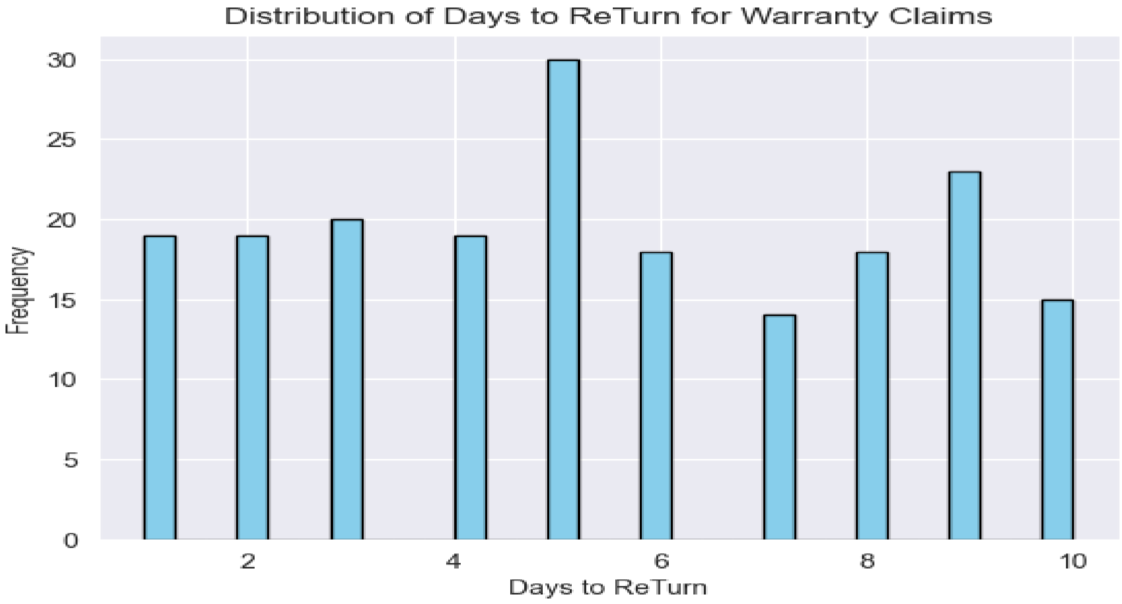
5. **Product Categories:** I’ve visualized the distribution of product categories. This can help understand which categories are most popular, I suggest we offer more product in those two categories like electronics and clothes and it preferred to offer products that really the customer need or highly demanded products



6. **Governorate Analysis:** I ’ve analyzed the average satisfaction score and number of Cairo and Giza and Alex taking the lead and it also oblivious because they are the most populated gouvernantes These urban centers naturally attract more attention due to their size and significance. so, our goal to make more market campaigns in the other governates and also try to target this population more because rural markets often remain untapped due to misconceptions or lack of targeted efforts. By reaching out to these areas, the company can expand their customer base and diversify revenue streams.

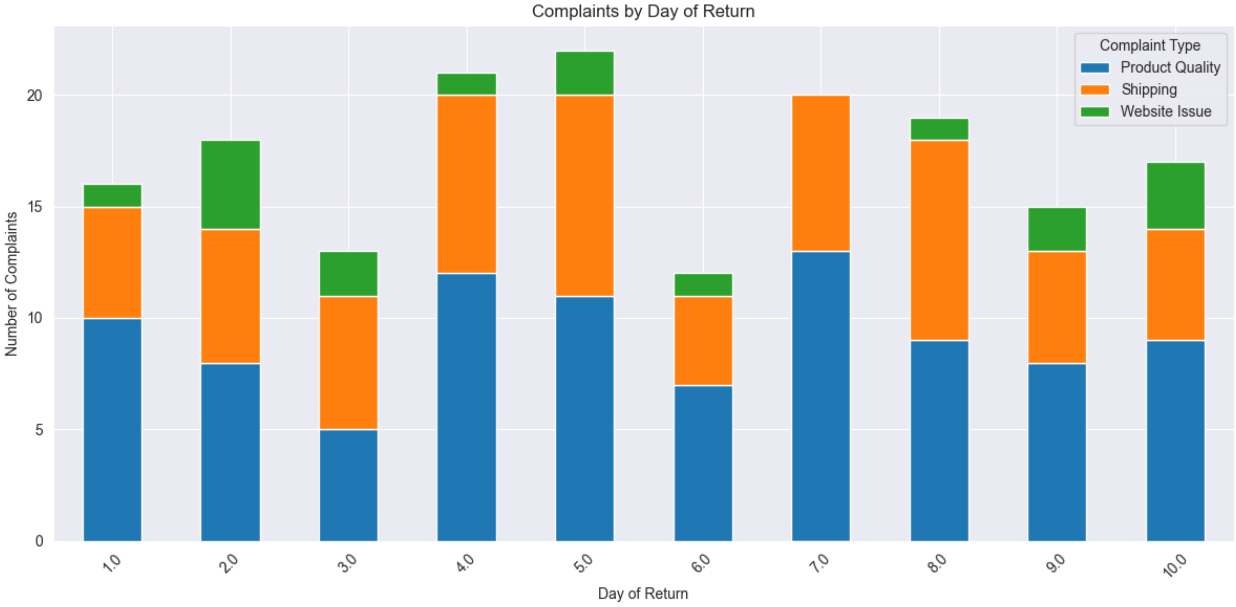


7. **Return:** the most common time taken for returns is 5 days, customers are generally prompt in initiating returns after receiving the product. Because Customers may be discovering issues with the product shortly after receiving it, leading to returns. Ensure that product descriptions and images are accurate and that the quality of the products meets or exceeds customer expectations. And also evaluate the packaging of our products to ensure that items are well-protected during shipping. Damage during shipping could lead to returns



hint I restarted the kernel and my whole data got randomized

the reason from this graph is to show most reason to see the product are being returned Shipping 16 Product Quality 15 Website Issue 14



Third Machine Learning:

First cluster analysis: This analysis provides insights into customer segments based on their purchasing behavior, profitability, and satisfaction levels, which can be used for targeted marketing strategies and personalized customer experiences.

Cluster	Recency	Frequency	Monetary	Quantity	Customer Profit	Total Profit BY Cust	Customer Satisfaction	Customers
0	147.6	2.03	579.60	10.80	47.05	241.76	4.59	51
1	299.73	1	203.67	4.991	18.64	89.706	4.55	340
2	113.93	1	192.78	4.820	18.51	85.67	4.9	356
3	209.53	1	796.5	5.615	49.65	271.25	4.91	200

Cluster 0: These customers have a relatively recent purchase history (average recency of 147.63 days), purchase items infrequently (average frequency of 2.04), and spend moderately (average monetary value of 579.61). They buy items in moderate quantities (average quantity of 10.80) and generate a moderate amount of profit per purchase (average customer profit of 47.06). The total profit generated by these customers is 241.76, and their average satisfaction score is 4.59. so, the goal to Target these customers with promotions or discounts to encourage more frequent purchases and larger order quantities.

Cluster 1: These customers have a less recent purchase history (average recency of 299.74 days), purchase items less frequently (average frequency of 1.00), and spend less (average monetary value of 203.68). They buy items in smaller quantities (average quantity of 4.99) and generate a lower amount of profit per purchase (average customer profit of 18.65). The total profit generated by these customers is 89.12, and their average satisfaction score is 4.55. so, the goal to Try to re-engage these customers with personalized offers or reminders to increase their frequency of purchase.

Cluster 2: These customers have a more recent purchase history (average recency of 113.93 days), purchase items less frequently (average frequency of 1.00), and spend less (average monetary value of 192.78). They buy items in smaller quantities (average quantity of 4.82) and generate a lower amount of profit per purchase (average customer profit of 18.51). The total profit generated by these customers is 85.65, and their average satisfaction score is 4.96.

So, the goal to Focus on improving the customer experience to maintain their high satisfaction level and encourage repeat purchases.

Cluster 3: These customers have a moderately recent purchase history (average recency of 209.53 days), purchase items less frequently (average frequency of 1.00), and spend more (average monetary value of 796.50). They buy items in moderate quantities (average quantity of 5.61) and generate a higher amount of profit per purchase (average customer profit of 49.65). The total profit generated by these customers is 271.25, and their average satisfaction score is 4.91.

Since those customers is high-value, so consider loyalty programs or exclusive offers to retain them and encourage them to spend more.

Second Random Forest Classifier Training for Customer Satisfaction Prediction:

The dataset was prepared for predicting customer satisfaction levels. First, categorical variables such as Complaint Type, Product Category, Product, Payment Method, and Governorate were encoded using Label Encoder to convert them into numerical values suitable for machine learning algorithms. The dataset was then split into training and testing sets with an 80/20 ratio using the train test split function. To ensure that the features were on the same scale, Standard Scaler was applied to standardize the data, making the mean of each feature 0 and the standard deviation 1. Finally, a Random Forest Classifier was trained on the standardized training data to predict Satisfaction Level, achieving an accuracy score of 1 on the test set (this score is normal to achieve if data is simulated).

I compared this model accuracy with the logistic regression model (0.925) which are actually good but the RF was better

Confusion Matrix:

	Predicted Not satisfied	Predicted satisfied
Actual Not satisfied	58	9
Actual satisfied	6	127

- 127 - These are the cases where the actual value was "Yes" and the model predicted "Yes" correctly.
- 58 - These are the cases where the actual value was "No" and the model predicted "No" correctly.
- 9 - These are the cases where the actual value was "No" but the model predicted "Yes" incorrectly.
- 6 - These are the cases where the actual value was "Yes" but the model predicted "No" incorrectly.

Now let’s go to the prediction Model using the RF

The Random Forest model achieved an accuracy of 100% on the test set. To demonstrate its application, the stat function was employed to predict customer satisfaction for a sample customer with the following characteristics: cost of item 800, shipping duration 2, complaint type 4, product category 2, specific product 5, product review 5, payment method 10, residing in governorate 200, purchasing a quantity of 3 items, customer profit of 1, and total profit by customer of 5. The model predicted that this customer is satisfied

Ok, what we can do with this prediction or this model? Firstly, we can Identify factors that contribute to lower satisfaction levels and take corrective actions to improve them. This could involve improving product quality, enhancing shipping processes

Secondly, we can identify customers who are likely to be dissatisfied and target them with specific marketing campaigns or offers to improve their experience and retain them as customers.

Third Recommendations systems:

The first: recommendation system is designed to provide personalized product recommendations to customers based on their past purchase behavior. Here's how it works:

1. **Input**: The system takes a customer ID as input, representing the customer for whom recommendations are sought.
2. **Finding Bought Products**: It first identifies the products that the customer has already purchased from the dataset.
3. **Calculating Similarity**: For each of these purchased products, the system calculates the similarity to all other products in the dataset. This similarity measure is typically based on how frequently two products are purchased together by different customers.
4. **Generating Recommendations**: Based on the similarity scores, the system identifies products that are similar to the ones the customer has already bought but has not yet purchased. These similar products are then recommended to the customer.
5. **Output**: The system provides a list of recommended products for the customer, which they may consider purchasing.

Benefits: Increasing the sales by suggesting relevant products, the system can help increase the likelihood of additional purchases by customers.

The second one Is similar but differs in the input in which I enter the product not the customer ID

Fourth conclusion:

The synthetic data simulation for Taager has provided valuable insights into customer behavior, product performance, and overall business operations. The study has identified key areas of focus such as product quality, shipping processes, and customer satisfaction. The application of machine learning models has further enhanced the understanding of customer segments and their purchasing behavior.

The research has also highlighted the potential of recommendation systems in enhancing the customer experience and driving sales. By providing personalized product recommendations, these systems can encourage customers to make additional purchases, thereby increasing revenue.

Moreover, the predictive models developed in this study can be instrumental in identifying customers who are likely to be dissatisfied. This allows for proactive measures to improve their experience and retain them as customers.

And also the findings from this study can guide strategic decisions and operational improvements for Taager. By addressing the identified issues and leveraging the insights gained, Taager can further optimize its platform, empower entrepreneurs, and continue to revolutionize social e-commerce in the MENA region. The lessons learned can also contribute to the broader understanding of social e-commerce dynamics, benefiting other players in the industry.

This research is a testament to the power of data analysis and machine learning in driving business growth and innovation. As the field continues to evolve, it will undoubtedly play an increasingly critical role in shaping the future of e-commerce.

Recommendations:

Based on the findings from the synthetic data simulation for Taager, the following recommendations are proposed:

1. **Product Quality**: Address the product quality issues identified in the complaint types. Consider dealing with alternative products that have higher quality, especially for the products that received the most complaints about their quality, such as mobile phones, hats, dresses, and boards.
2. **Shipping Process**: Evaluate the shipping process, especially for shipping type 2 which received the most complaints. Consider improving the packaging of products to ensure that items are well-protected during shipping. Damage during shipping could lead to returns and dissatisfaction.
3. **Payment Methods**: Continue to offer popular payment methods like ‘Valu’ and ‘Cib bank’. Consider introducing more payment options to cater to a wider range of customer preferences.
4. **Product Categories**: Offer more products in popular categories like electronics and clothes. Ensure that the products offered meet customer needs and are in high demand.
5. **Governorate Analysis**: Conduct more marketing campaigns in other governorates apart from Cairo, Giza, and Alex. Rural markets often remain untapped due to misconceptions or lack of targeted efforts. By reaching out to these areas, the company can expand its customer base and diversify revenue streams.
6. **Returns**: Ensure that product descriptions and images are accurate and that the quality of the products meets or exceeds customer expectations. This can help reduce the number of returns.
7. **Machine Learning Models**: Leverage the predictive models developed in this study to identify customers who are likely to be dissatisfied. Target them with specific marketing campaigns or offers to improve their experience and retain them as customers. Also there s so many models was needed to be trained in my opinion but igot capped due to the simulation such as multiple regression and multiple time series model (LSTM,GRU,CNN ,ARIMA)