**CUSTOMER SEGMENTATION**

We first sought to understand the problem and gained insights into the industry. We researched how best to organize customer segmentation. Using information from our raw dataset, we generated new insights to improve our analysis. We effectively dealt with outliers and missing values. We examined correlations between variables and used the PCA algorithm to transform our data into three dimensions. Employing the KMeans machine learning algorithm, we created 5 different clusters. We analyzed relationships within these clusters and conducted a market campaign accordingly.

**EXPLORATORY DATA ANALYSİS:**

1)First of all, we removed the variables from our dataset that we could not obtain useful information at the beginning of the project.

-Customer ID and Customer Name

2) We have created more meaningful variables by transforming the variables that we will use more meaningfully. Or we converted them into binary form.

Customer\_gender = if male 1 , if female 0

Customer\_birth\_date = We calculated the age of the customer birthdated person and changed it to age.

Loyalty\_card\_number = By giving 1 to those who have a loyalty card number and 0 to those who do not have a loyalty card number, we have obtained information according to whether there is a loyalty card number or not.

Latitude and longitude=By calculating the coordinates from the coordinates of the centre of Lisbon using the Euclidean method. We calculated the distance of the customer to the city centre.

year\_first\_transaction=By subtracting today's date from year\_first\_transaction, we get how long the customer has actually been a customer.

Typical Hour=typical\_hours to extract meaningful information about people's work hours according to their work schedule time\_slots = {

"early\_morning": range(6, 9),

"morning": range(9, 13),

"afternoon": range(13, 18),

"evening": range(18, 22),

"night": range(22, 24)

grouped as.

**HOW WE FİLLED THE EMPTY VALUES.**

* **kids\_home 524**
* **teens\_home 1021**
* **number\_complaints 654**
* **distinct\_stores\_visited 1309**
* **typical\_hour 1745**
* **lifetime\_spend\_vegetables 873**
* **lifetime\_spend\_fish 1309**

There were empty values in our variables above, and while dealing with these empty values, we used the missigno library to check if there was a correlation between them. And since there was no correlation between them above the threshold value, we decided to fill the empty values.

We used customer\_basket to fill some empty values in life\_spend but we were not successful.

A group of blue squares with black text

Description automatically generated with medium confidence

While filling our empty values, we looked at our distributions, some of them had anomalies and outlierz, accordingly, we estimated the empty values using RandomRegressor or filled them with empty values.

**How we dealt with outliers:**

While dealing with outliers, we also investigated the sector we are dealing with and realised that there are values that appear to be outliers but that we can make sense of.

A map of a person's face

Description automatically generated A graph of a long line

Description automatically generated with medium confidenceFor example, the distance of the customers may appear as an outlier, but they are data that we can obtain information as we will explain later. It may allow us to segment them as customers coming from a long distance.

Again, by looking at the distributions of the variables with box\_plot, we suppressed some of them to the upper and lower levels and some of them were not touched at all as in distance. We also examined in detail how much of our total data the outliers constitute.

A graph of a fish

Description automatically generated

We detected an anomaly in life spend fish and decided to leave it like this as we thought we could make sense out of it.

**RELATİONSHİPS BETWEEN VARİABLES**

A screenshot of a computer

Description automatically generated

Of course, the best visualisation tool to see the relationship between variables is the correlation matrix. We will now look at the high coefficients between the correlations. And we'll check for correlations.

A graph with blue lines

Description automatically generated A graph with blue dots

Description automatically generatedA graph with blue dots

Description automatically generated  
 A graph with blue dots

Description automatically generated A graph with blue dots

Description automatically generated

There is a -0.28 correlation between the money spent on pet food and the money spent on vegetables. This may indicate that people who buy pet food are slightly less likely to consume vegetables.

The value in the matrix is 0.83, which means that the further away from the city, the higher the expenditure on fish shopping.

As electronic shopping increases, so does meat shopping. We have said the same thing about video games before, maybe because there is an important relationship between video games and electronics.

The value in the matrix is 0.83, which is a large value, and when we look at the graph, there is a high relationship between video game expenditures and meat expenditures.

The coefficient between them is -0.36, which means that those who buy electronic products do not prefer promotional products. Also consider the link between electrical products and meat shopping and video games.

A graph of a graph with numbers and lines

Description automatically generated with medium confidence

There is a -0.26 negative correlation between pet food and having children at home. This could mean that as the number of children in the household increases, spending on pet food decreases. In reference to the positive relationship between kids and teens, it can be said that as the number of teens at home increases, spending on pet food decreases.

A graph with blue lines

Description automatically generated

As previously mentioned, we have already grouped the hours, but here we see that there is significantly more shopping between 9 AM - 1 PM and right after work hours.

**MODEL SELECTİON AND PARAMETER TUNİNG.**

A graph with blue lines

Description automatically generatedA graph with a blue line and green line

Description automatically generated

After performing PCA to train the model and finding the optimal number of components to be 3, we conducted a 3-dimensional reduction. Subsequently, using the KElbow method with KMeans algorithm, we determined that there should be 5 clusters

**CUSTOMER SEGMENTATİON**

**We will now analyse all the clusters and their main characteristics through graphs and statistics.**

**PRED CLUSTER 2 -)**

The most basic characteristics of the customers we assigned as **pred cluster 2**

-They live far from the city centre

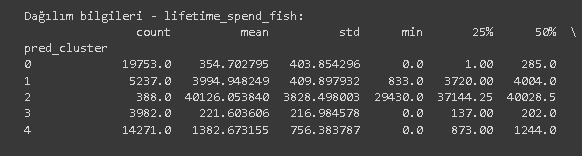
-They spend too much money shopping for fish

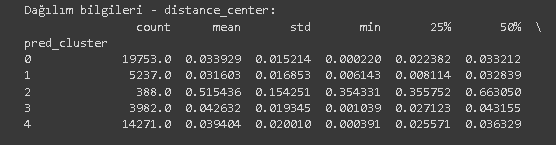
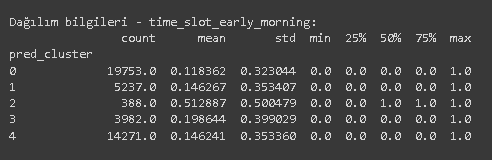
-And they shop mostly in the early morning.

-And there are almost no kids and teens in their homes.

We can call this cluster **Urban Fishmongers.**

A screenshot of a computer screen

Description automatically generated

A graph with a bar chart

Description automatically generated with medium confidence A graph with a blue bar

Description automatically generatedA diagram of a fish

Description automatically generated with medium confidence

**PRED CLUSTER 0=( Diverse and Budget-Conscious Shopper**"**)**

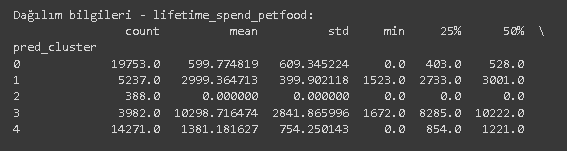
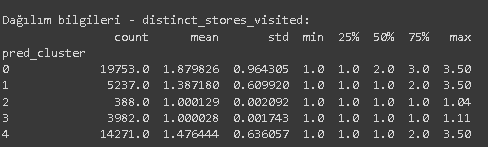
 Low spending on pet food indicates limited expenditure or infrequent purchases.

 Frequently visiting different stores suggests a preference for variety and comparison shopping.

 Preferring discounted products reflects a budget-conscious approach to shopping.

*  Placing importance on vegetable purchases highlights a commitment to healthy eating habits.
* This cluster can be called "**Diverse and Budget-Conscious Shopper**"

A screenshot of a computer

Description automatically generated  

A graph with blue squares

Description automatically generatedA graph of a graph

Description automatically generated with medium confidenceA screen shot of a computer

Description automatically generated

A graph with blue squares and black lines

Description automatically generated

CLUSTER 1 = **Tech and Wellness Enthusiasts**

Here are the characteristics of this customer group listed as bullet points:

* Typically consists of young individuals with a strong interest in technology.
* They show a significant interest in video games, which plays a crucial role in their leisure and recreational activities.
* They take great care in personal grooming; investing in hygiene products and health supplements.
* They prioritize meat consumption for healthy eating and prefer various meat products.
* They invest in technology products to simplify their daily lives and keep up with innovative products.

A graph with blue squares and black text

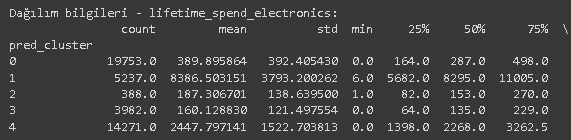
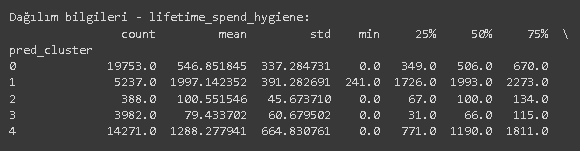
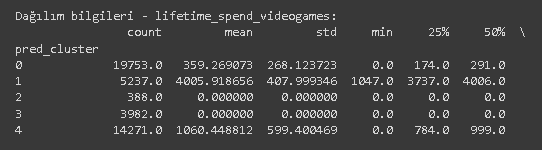
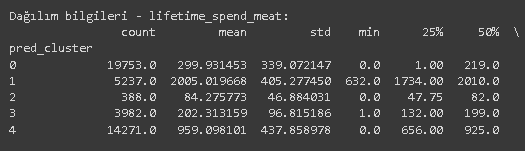
Description automatically generated A graph with blue squares and black text

Description automatically generated

A graph of a graph

Description automatically generated with medium confidenceA graph of a diagram

Description automatically generated with medium confidence

**CLUSTER3 Pet-Centric Wellness Shoppers**

Here are the characteristics of this customer group listed as bullet points :

* They visit different markets infrequently.
* They prefer non-alcoholic beverages.
* They spend significantly on pet food.
* They generally have few or no children at home.
* They are predominantly women.

This customer group likely consists of individuals who prioritize their pets' needs and embrace a healthy lifestyle.

**A graph of a graph

Description automatically generated with medium confidence A graph with blue squares

Description automatically generated**

**A graph of a bar chart

Description automatically generated with medium confidence A graph of a bar chart

Description automatically generated**

A screenshot of a computer screen

Description automatically generated

**CLUSTER 4 Big Family Shoppers**

Here are the characteristics of Cluster 4 customer segment listed as bullet points:

* They have a large family structure at home.
* They spend less on pet food.
* They make average-level expenditures in general shopping.
* They frequently purchase a variety of distinct products.

This customer segment not only consists of large families with moderate spending habits but also shows a tendency to purchase a wide range of different products.

**A graph with a blue square and black lines

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence**

A screenshot of a graph

Description automatically generated**A graph with a blue square and black squares

Description automatically generated**

In addition, the distribution of our clusters is such that it is obvious that cluster 2 (**Urban Fishmongers**)forms a very different cluster from the others.

A graph showing different colored shapes

Description automatically generated

Number of customers in each cluster.

Cluster 0: 19753

Cluster 1: 5273

Cluster 2: 388

Cluster 3: 3982

Cluster 4: 14271

MARKET CAMPAIGNS

\*For Cluster 2 (**Urban Fishmongers**):

They typically do their fish shopping in the morning, and there's a 0.37 connection between fish shopping and groceries. To support this connection:

1. Early morning fish shoppers receive a 15% discount on groceries.
2. To sell more fish, there's a 25% discount on every $15 spent on fish shopping in the early morning.

\*For Cluster 1 (**Tech and Wellness Enthusiasts**):

We've already discussed how closely related electronics, video games, and personal hygiene are, and now it's campaign time related to these.

-Get a 10% bonus points on purchases of $100 or more on electronics, video games, and personal hygiene products.

(This is because there's a tendency among these customers to purchase discounted items, indicated by a correlation of -0.36.)

\*For Cluster3 (**Pet-Centric Wellness Shoppers**):

They spend a lot on pet food and prefer non-alcoholic beverages. Additionally, their vegetable purchases are directly proportional to these.

To perhaps encourage pet owners to buy vegetables and positively influence this behavior, we can link these two negative elements.

Get a $5 cash gift on vegetable purchases for pet food expenditures of $20 or more.

\*For Cluster 4(**Big Family Shoppers**):

We know there are a lot of children in this cluster, and there's a 0.22 correlation with the typical hour at home being influenced by kids and teens. Additionally, although spending is average on almost everything, having children also affects meat consumption, as seen from the correlation table.Therefore, we can encourage meat shopping late at night.

20% discount on meat products after 7:00 PM.

\*For Cluster 0(**Diverse and Budget-Conscious Shopper**):

We observe that customers in this cluster visit many stores and tend to purchase discounted products. This suggests that targeted campaigns could be launched to attract new customers who actively seek discounts.

"For first-time membership card holders, a $15 coupon valid for purchases of $50 or more."

**CONCLUSİON:**

We conducted customer segmentation on our market data to best tailor campaigns to customers, enhance their shopping experience, and garner favor towards our company. We prepared the data for analysis through exploratory data analysis, addressing missing values, identifying empty variables, and generating new variables. Using data visualization methods, we examined relationships between variables. PCA analysis and KMeans clustering resulted in 5 clusters, which were analyzed using statistical techniques and visualization methods. A main theme was identified for each cluster, and market campaigns were developed accordingly. The relationships and connections within the clusters yielded positive results consistent with the initial correlation matrix. Our project has been successful to a significant extent.

Kemal Can Özmetli 20231931

Artem Khomytskyi; 20221686

Tymofii Kuzmenko; 20221690