## **Topic**

**Research on the effectiveness of advertising campaigns in the field of marketing by analyzing customer behavior and marketing metrics**

## Project description

My work will focus on researching data on user interaction with advertising campaigns, their demographics, and behavior. Analyzing the demographic characteristics of users and their income helps to understand their needs better and, accordingly, their behavior. In turn, the research on behavior and interaction with advertising campaigns makes it possible to choose the best marketing strategies for the highest efficiency in working with users.

The work will include segmentation of users by gender, age, and income groups, analysis of the presence or absence of influence of the budget spent on an advertising campaign on its performance indicators, creation of a customer loyalty rating, and research of the behavior of the most loyal of them, as well as loyalty segmentation for use in further advertising campaigns.

The work aims not only to investigate the above aspects but also to provide business recommendations on the best channels and types of advertising campaigns to achieve the highest possible marketing efficiency.

## Technologies stack

**Power Query:** cleaning data and determining the correct data types for each variable;

**Excel:** descriptive statistics, box plots, correlation analysis;

**Google BigQuery (SQL):** preparing data for building visualizations, creating age and income categories, creating a ranking list, and aggregating metrics by channels and types of campaigns;

**Power BI:** data visualization;

**Jupyter Notebook (Python):** segmentation of customers by loyalty for its further use in marketing strategy.

## Dataset selection

The Kaggle platform was chosen for the selection of the dataset due to its versatility and variety of materials. The dataset selected for the project:

Rabie El Kharoua. (2024). 📈 Predict Conversion in Digital Marketing Dataset [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/8742946>

I liked the dataset for the volume of data and the possibility of their multifaceted analysis.

Such databases as [BigQuery public datasets](https://cloud.google.com/bigquery/public-data) and [*UCI Machine Learning Repository*](https://archive.ics.uci.edu/) were also investigated, but they were not chosen because of their rather specific subject matter.

### Dataset description

**Demographic Information**

* **CustomerID**: Unique identifier for each customer.
* **Age**: Age of the customer.
* **Gender**: Gender of the customer (Male/Female).
* **Income**: Annual income of the customer in USD.

**Marketing-specific Variables**

* **CampaignChannel**: The channel through which the marketing campaign is delivered (Email, Social Media, SEO, PPC, Referral).
* **CampaignType**: Type of the marketing campaign (Awareness, Consideration, Conversion, Retention).
* **AdSpend**: Amount spent on the marketing campaign in USD.
* **ClickThroughRate**: Rate at which customers click on the marketing content.
* **ConversionRate**: Rate at which clicks convert to desired actions (e.g., purchases).

**Customer Engagement Variables**

* **WebsiteVisits**: Number of visits to the website.
* **PagesPerVisit**: Average number of pages visited per session.
* **TimeOnSite**: Average time spent on the website per visit (in minutes).
* **EmailOpens**: Number of times marketing emails were opened.
* **EmailClicks**: Number of times links in marketing emails were clicked.

**Historical Data**

* **PreviousPurchases**: Number of previous purchases made by the customer.
* **LoyaltyPoints**: Number of loyalty points accumulated by the customer.

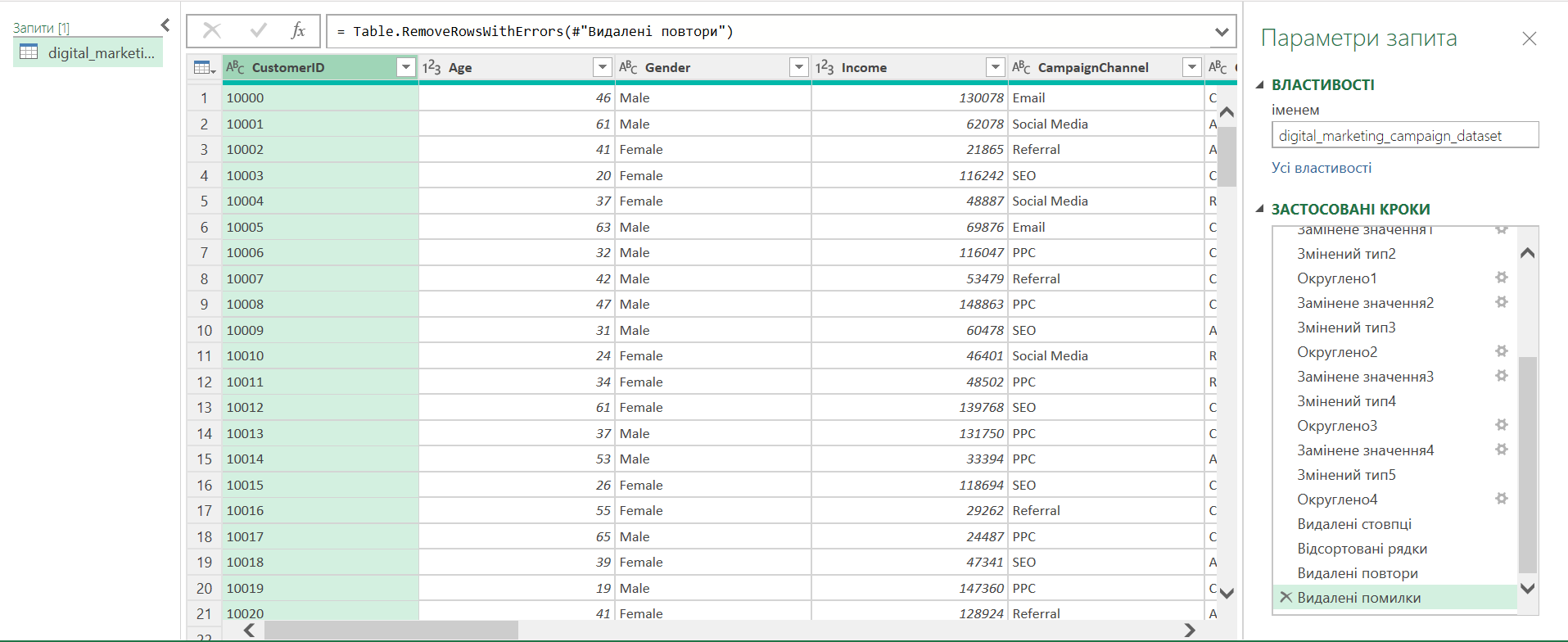
**Target Variable**

* **Conversion**: Binary variable indicating whether the customer converted (1) or not (0).

## Plan

1. Data cleaning and transformation, preparation for further analysis in Power Query.
2. Using descriptive statistics to better understand data, and a data analysis package to perform correlation analysis in Excel.
3. Loading data into Google BigQuery.
4. Prepare data for visualization by adding new columns and aggregation.
5. Load data and create visualizations in Power BI.
6. Performing segmentation in Python.

### Data cleaning



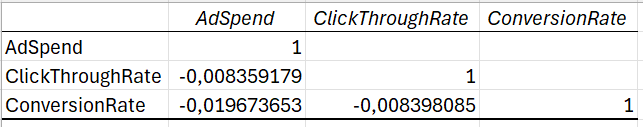
In the process of preparing and cleaning the data, the correct data types were determined in the columns, empty entries and errors were eliminated. The results showed the purity of the dataset because the number of records did not change after that.

### Description Statistics

<https://docs.google.com/spreadsheets/d/1y8BaWVfx7fVvt-8Hk62EePpUo2WdJTHL/edit?usp=sharing&ouid=104763776962176246792&rtpof=true&sd=true> (since the work was created in Excel, Google Sheets does not display some elements, including box plots).

Descriptive statistics were generated for each numeric field and box plots were constructed to identify outliers in the data, if present. The statistics helped later define the framework for the user groups, and the box plots showed the absence of outliers in the data.

### Correlation Analysis



Correlation analysis showed a very weak negative correlation between the budget spent on advertising campaigns and their effectiveness (the main metrics of CR and CTR). So, in this case, the amount of money spent does not affect the efficiency.

### SQL

SELECT

CustomerID,

Gender,

Age,

--Creating age categories

CASE

WHEN Age >= 18 AND Age < 31 THEN "young"

WHEN Age >=31 AND Age < 51 THEN "middle"

ELSE "senior"

END age\_category,

Income,

--Creating income categories

CASE

WHEN Income >= 20000 AND Income < 50000 THEN "low"

WHEN Income >= 50000 AND Income < 100000 THEN "moderate"

ELSE "high"

END income\_category,

WebsiteVisits,

PagesPerVisit,

TimeOnSite,

EmailOpens,

EmailClicks,

--Counting personal activity rate (shows the ratio of email clicks to email opens)

ROUND(CASE

WHEN EmailOpens = 0 THEN 0 --considering the presence of 0 values in the divider

ELSE (EmailClicks / EmailOpens) \* 100

END, 2) AS personal\_activity\_rate,

PreviousPurchases,

LoyaltyPoints,

--ranking the customers by loyalty points without gaps in rating values caused by the same points values

DENSE\_RANK() OVER(ORDER BY LoyaltyPoints DESC) as loyalty\_rating,

Conversion

FROM `pet-project-428819.marketing\_data.data`

ORDER BY loyalty\_rating

**Results**: <https://drive.google.com/file/d/1Zv8aTfffYoIdsp02Z4oGy898FBjY2-4o/view?usp=sharing>

SELECT

CampaignChannel,

--counting average spendings on the campaign channel

ROUND(AVG(AdSpend), 2) as avg\_ad\_spend,

--counting average CTR for each channel

ROUND(AVG(ClickThroughRate), 2) as avg\_ctr,

--counting average CR for each channel

ROUND(AVG(ConversionRate), 2) as avg\_cr

FROM `pet-project-428819.marketing\_data.data`

GROUP BY CampaignChannel

**Results:** <https://drive.google.com/file/d/1xvJ5x9_mhpdWYVi4eXx1NA-nfljD_zx5/view?usp=sharing>

SELECT

CampaignType,

--counting average spendings on the campaign type

ROUND(AVG(AdSpend), 2) as avg\_ad\_spend,

--counting average CTR for each type

ROUND(AVG(ClickThroughRate), 2) as avg\_ctr,

--counting average CR for each type

ROUND(AVG(ConversionRate), 2) as avg\_cr

FROM `pet-project-428819.marketing\_data.data`

GROUP BY CampaignType

**Results**: <https://drive.google.com/file/d/1B_ZZFPcBBwPCqFNZEQdIDbBnEq43LJpB/view?usp=sharing>

SELECT

--coverted users

SUM(Conversion) as total\_conversions,

--not converted users

COUNT(CustomerID) as total\_customers,

--full ratio of converted users

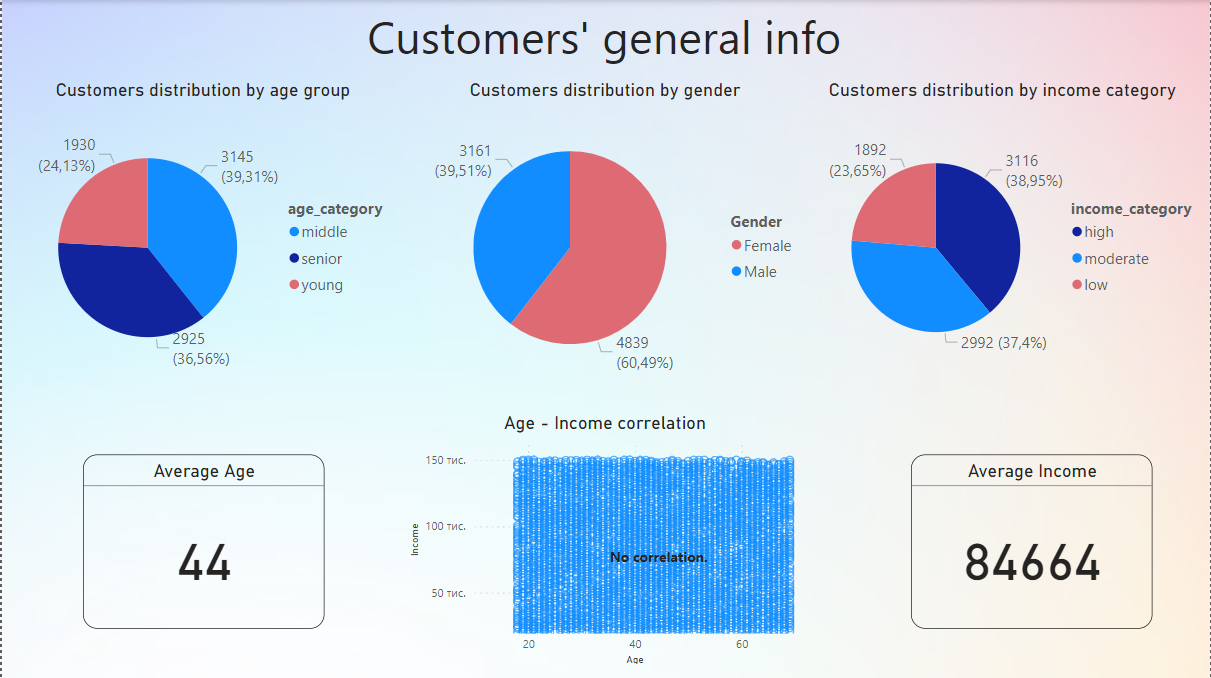
ROUND(SUM(Conversion)/(COUNT(CustomerID)), 2) as users\_conversion\_rate

FROM `pet-project-428819.marketing\_data.data`

**Results**: <https://drive.google.com/file/d/1McUrVEyjYPy4PV9KFE7elf-h_xn6I-ue/view?usp=sharing>

### Power BI

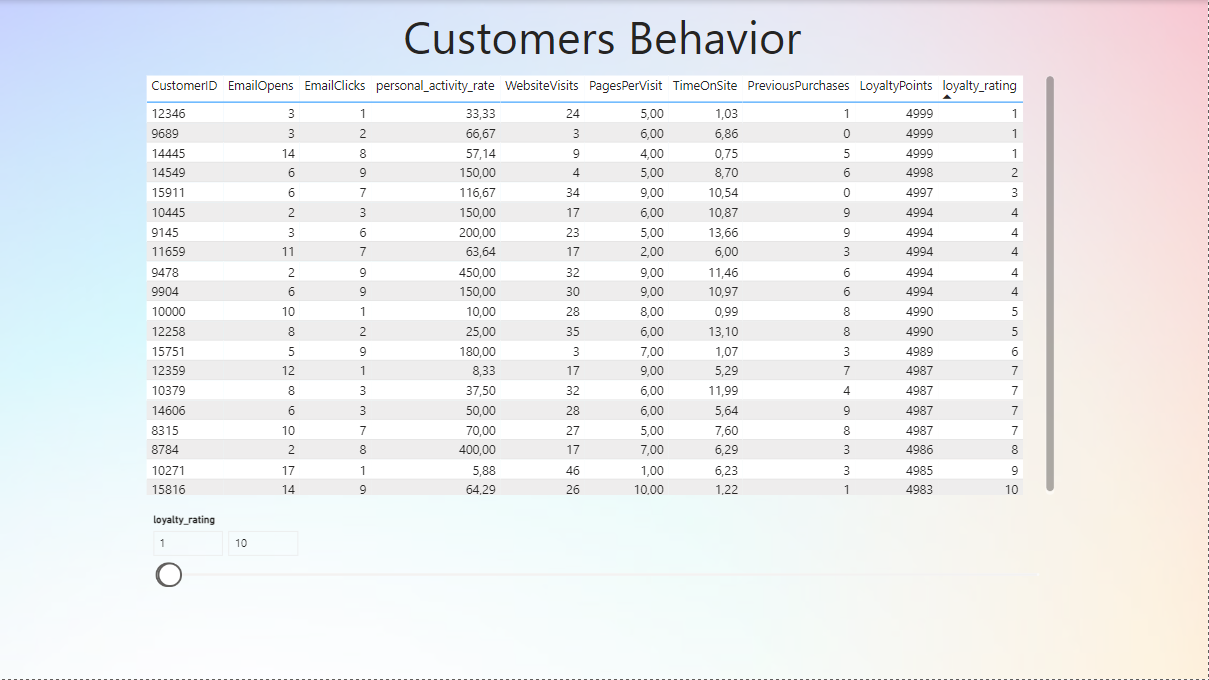
<https://drive.google.com/file/d/17Sh7PXEeHEK9kzAcRgOx6lAY1CFrlcLA/view?usp=sharing>



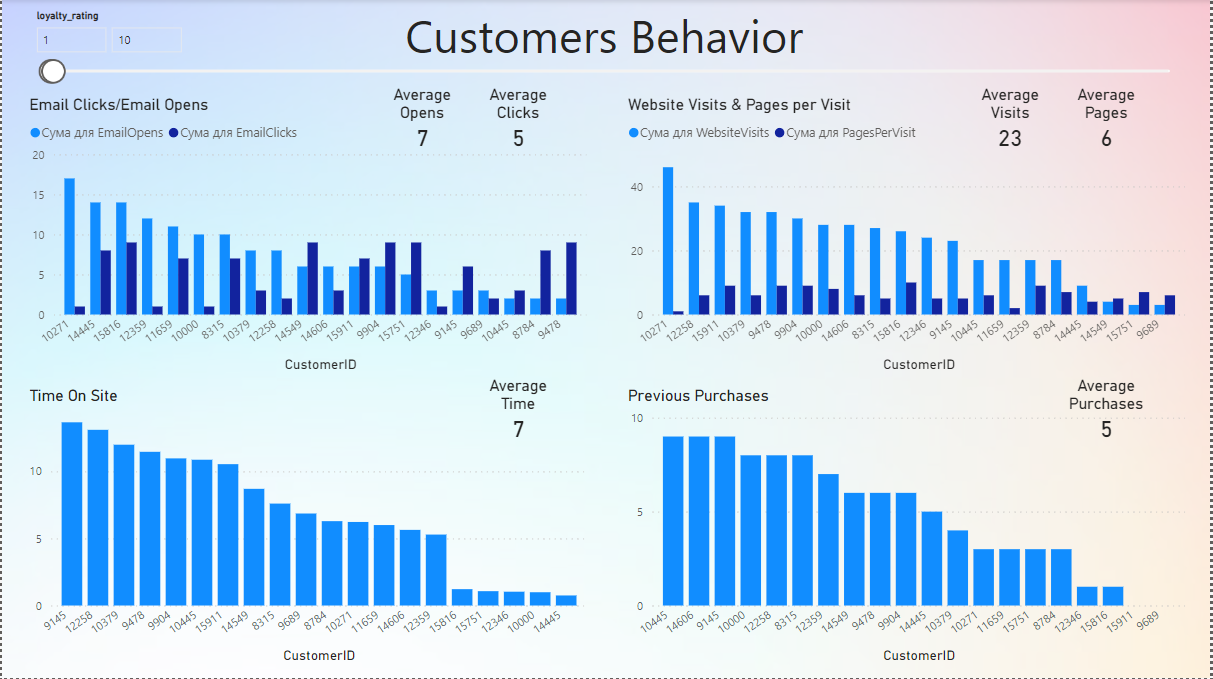
The first dashboard displays general information about users.

* By age, the largest share of users is the middle category (31-50 years old), in second place are senior customers (51-69 years old), and the smallest group of users is the young group (18-30 years old).
* By gender, the majority of our users are women (60%).
* In terms of annual income, high-income users (above $100,000) dominate, middle-income users ($50,000-100,000) have almost the same share and only 24% are low-income users ($20,000-50,000).
* The average age of all users is 44 years.
* The average income is $84,664.
* There is no correlation between age and income.

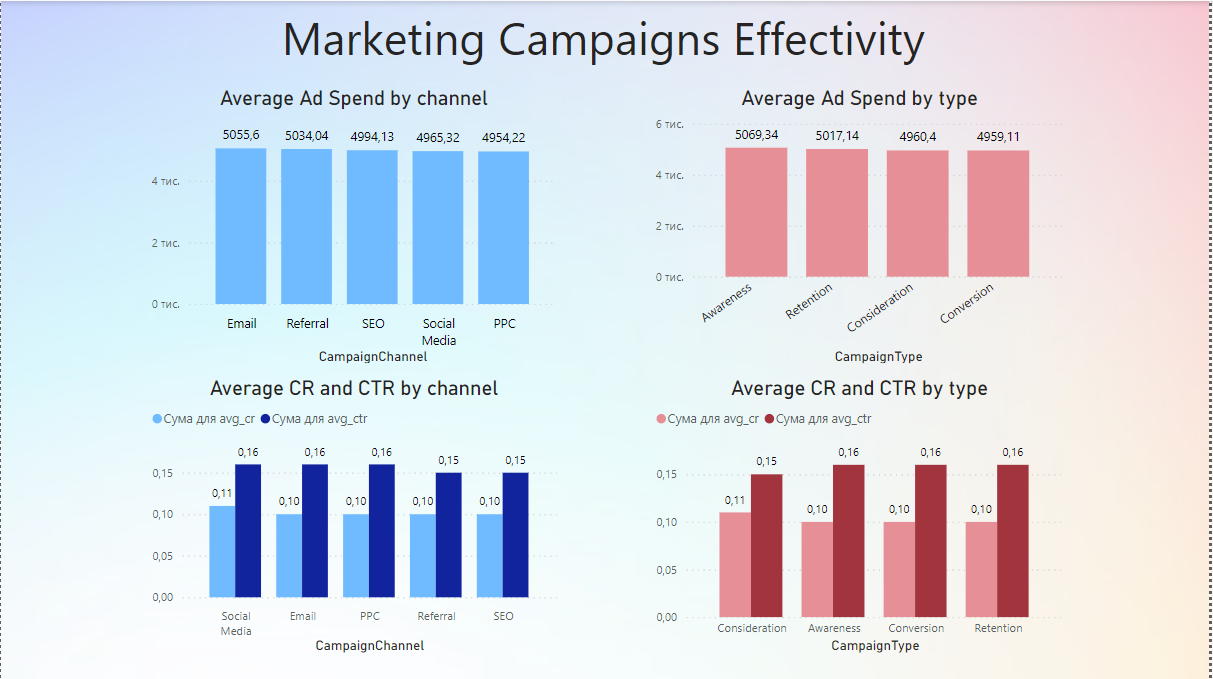
From the given data, it can be concluded that when preparing advertising campaigns, it is necessary to focus on the interests of older people, some of them can focus on the female audience, and it is also possible to advertise expensive or medium-priced products, because the audience is quite well-off.



On the second dashboard, you can view all information on customer behavior in a convenient format by filtering by user rating, as well as single out users with the required rating.



On the third dashboard, you can view a visual display of user behavior by filtering by rating. From the first plot, it can be seen that loyal users often return to emails and follow links again. The average number of open letters among TOP-10 loyal users is 7, and the average number of clicks is 5. The average number of site visits per user is 23 times, and the average number of viewed pages per visit is 6. The average amount of time spent on the site is 7 minutes, and the average number of previous purchases is 5 items.

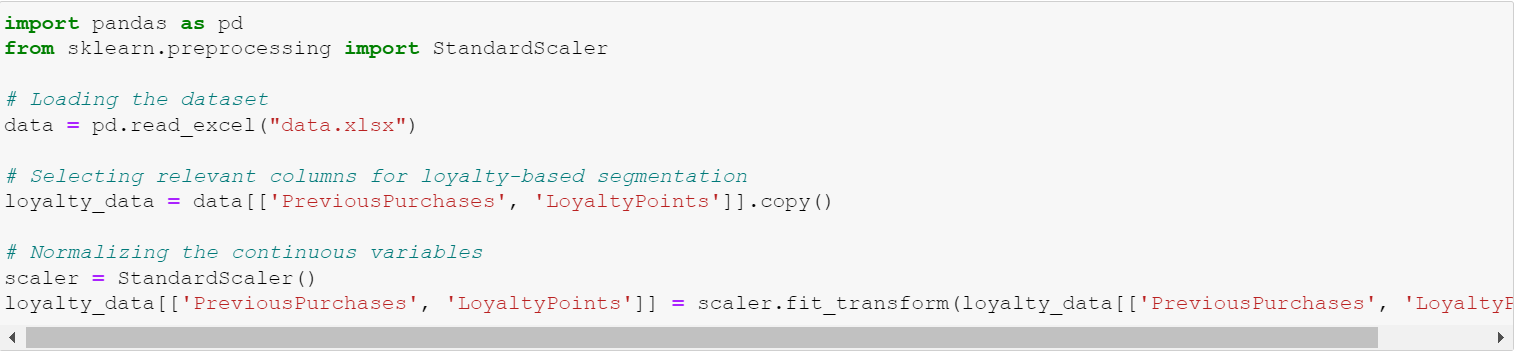


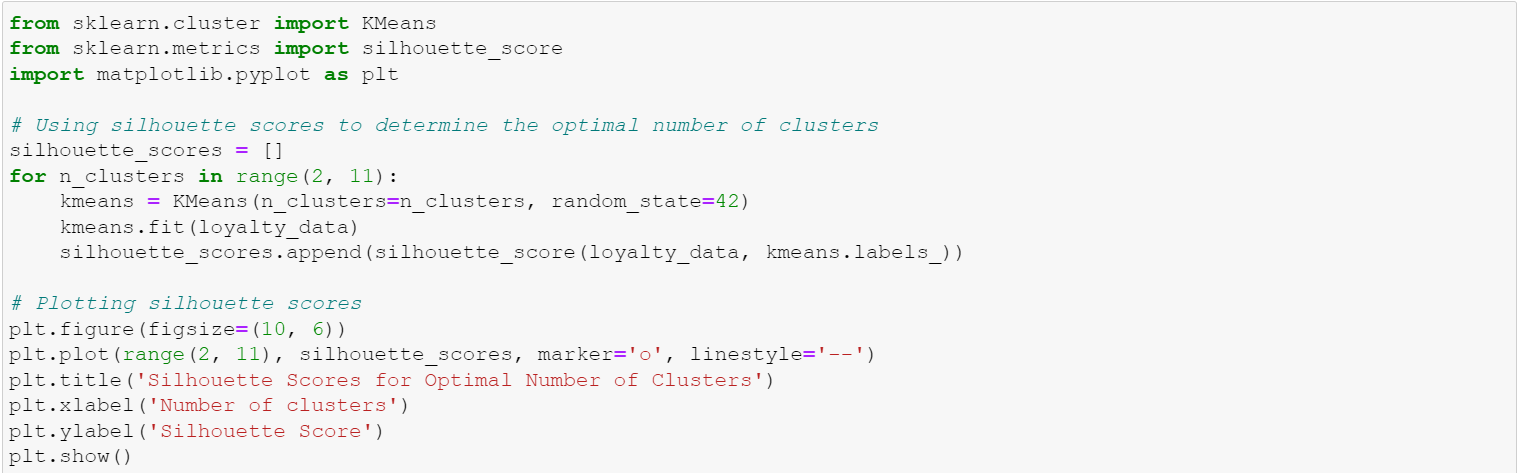
In the last dashboard, you can see the effectiveness of advertising channels and campaign types according to the main marketing metrics Click-Through Rate (CTR) and Conversion Rate (CR).

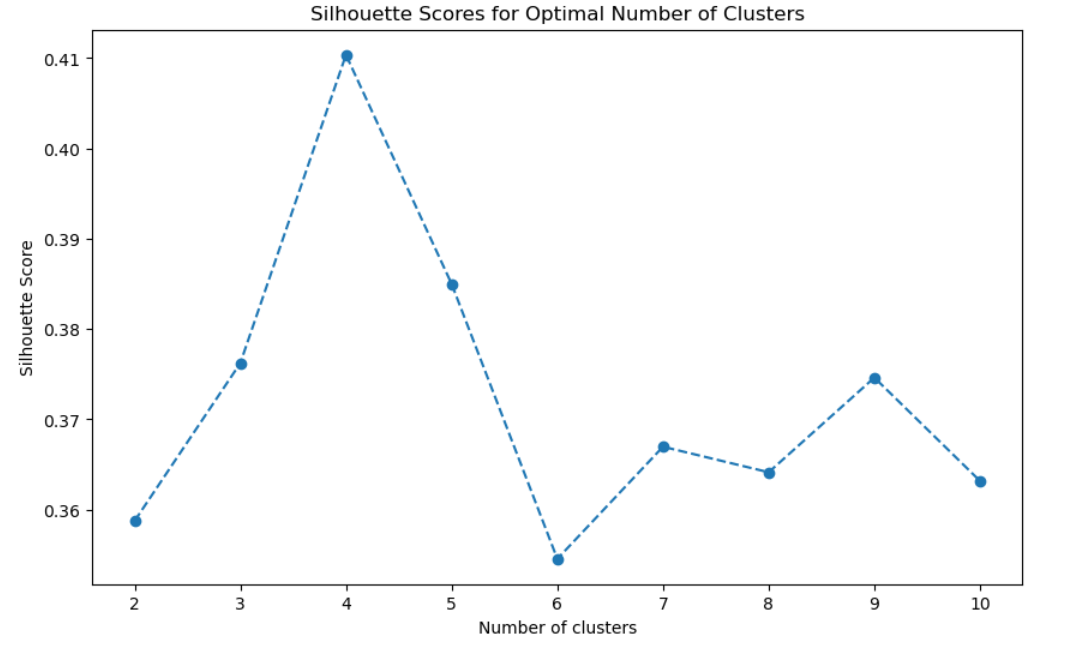
To conclude, the best channels for advertising campaigns in terms of cost and effectiveness are Social Media (highest effectiveness and low cost) and PPC (lowest cost and high CR), and types - Conversion (lowest cost and high CR) and Consideration (highest CTR and low cost).

### Loyalty-based segmentation

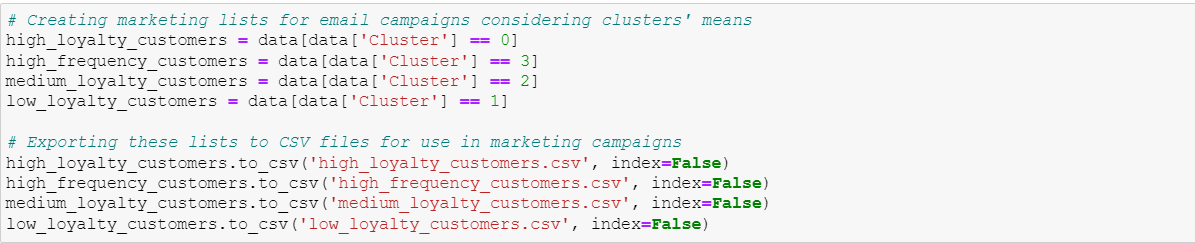
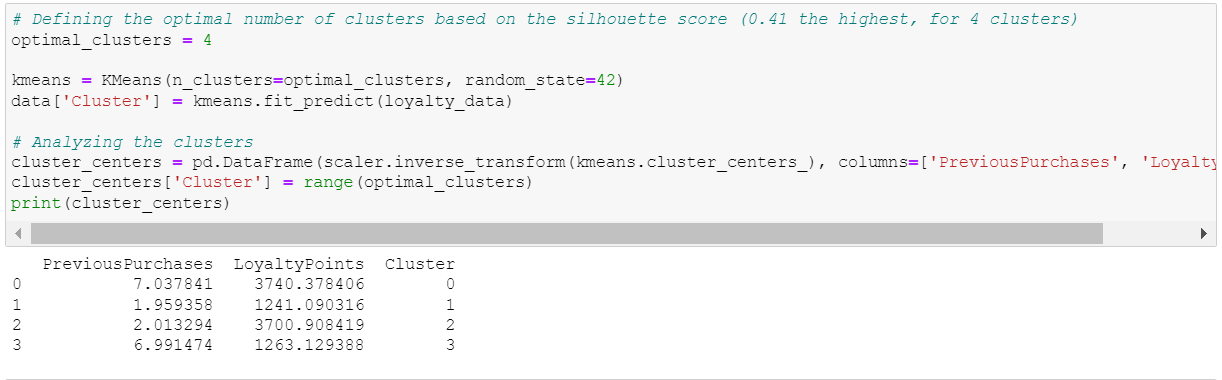
The KMeans method was chosen for segmentation because it is quite universal and one of the most popular and effective. The method perfectly copes with large volumes of data, dividing them into clusters from values ​​that most closely match the average of this cluster, rather than another.







Segmentation into 4 clusters was the most effective for this dataset because this option has the highest silhouette score. Therefore, further distribution of data into 4 clusters (segments):



У результаті сегментації було виділено 4 сегменти:

1. High loyalty customers: high level of loyalty points and many previous purchases;  
   A special loyalty system with discounts and special offers can be implemented for this segment.
2. High-frequency customers: many previous purchases, but not high values ​​of loyalty points;  
   For this segment, it is advisable to send offers with discounts often.
3. Medium loyalty customers: average values ​​of loyalty points and previous purchases;

Periodically send special offers and discounts.

1. Low loyalty customers: low values ​​of loyalty points and previous purchases;

Discounts on the first purchase after a long lull, promotions of the type 1+1 = 2.

## Conclusions

* The main audience of the business is people of middle and older age (from 30 years old) with average and high levels of annual income (from $50,000), more than half of them are women.
* The budget spent on advertising does not affect its effectiveness.
* There is no correlation between the income level of users and their age.
* Not all users from the TOP-10 by loyalty rating demonstrate high activity in terms of clicks, visited pages, time spent on the site, etc. Maybe they just made an expensive purchase that earned them a lot of points.
* About the same amount of money was spent on all channels and types of advertising cases.
* All campaigns showed almost the same performance in terms of CR and CTR. However, the best channels for ad campaigns by cost and performance are Social Media (highest performance and low cost) and PPC (lowest cost and high CR), and types are Conversion (lowest cost and high CR) and Consideration (highest CTR and low cost).
* Users can be divided into 4 loyalty segments to diversify advertising success.
* The business has a high user conversion rate (88%).

## Recommendations for business

* Creating advertising campaigns, focus on the interests and needs of the main audience (people of middle and older age (from 30 years old) with an average and high level of annual income (from $50,000), more than half of them are women).
* Try to improve the quality of advertising campaigns through Social Media by increasing the budget and analyze whether this will increase efficiency because the channel is promising in terms of indicators.
* Check the quality of all advertising campaigns and improve it to increase the performance indicators, because they are actually quite low.
* Use loyalty segmentation to create more diversified and user-targeted advertising campaigns to maintain or increase customer loyalty.