Cornell University

Quality analysis of online education advertising delivery channels

Yunhan Liu(yl3287) **Jiawei Gao(jg2352)**

I. Introduction

This case is an investigation of three months' worth of advertisement delivery data from a leading online education provider. Advertising is a crucial tool for Internet businesses to increase users, foster growth, and communicate ideals. A business might gain significantly by running high-quality advertisements on channels, yet doing the contrary could result in losses or even a disaster. Furthermore, businesses can discover the channels where their target audience is present and tailor their targeting strategies accordingly by studying the ad delivery channels. Additionally, it enables them to better their overall ad performance by understanding which channels are driving the most interaction, clicks, and conversions. Another significant benefit is that it aids companies in streamlining their advertising campaigns to boost return on investment (ROI). Therefore, a key component of top-level decision-making is the analysis of advertising delivery channel quality by data analysts.

In this instance, we will classify advertising channels with an unsupervised learning algorithm — K-Means and define their main characteristics using three months' worth of data on advertising delivery in order to inform further business decisions.

II. Data Exploration

The dataset is downloaded from Kaggle Competition (https://www.kaggle.com/datasets/ruizema123/advertising-channel-analysis). 13 variables and 900 rows make up the entire dataset. The variables can be classified into three categories, including 2 Channels variable (Channel_ID: Identity number of each Channel; Daily_UV: The number of unique visitors look through the website each day), 5 user behavior variables (Average_registration_rate: The ratio of the number of people registered to the daily unique visitors; Average_search_volumn: The average number of searches per unique visitor per day; Visit_depth: The number of pages a visitor views on a website during a single session; Average_session_duration: The average length of time that visitors spend on a website during a single session; Order_conversion_rate: The percentage of visitors who complete a purchase or transaction on a website, out of the total number of visitors to the website), and 6 advertising attributes variables (Total_advertising_duration: The length of time that an advertising to be exposure to the public; Material_type: Material type of the ads, such video, images, or text; Ad_type: Format or style of an advertisement, such as display ads, banner, tips etc; Cooperation_method: Cooperation method based on the performance of the advertising, such as CPC, ROI etc; Ad_size: Physical dimensions; Ad_selling_point: The key content topic of the advertising, such as the employment, the promotion, the low price etc).

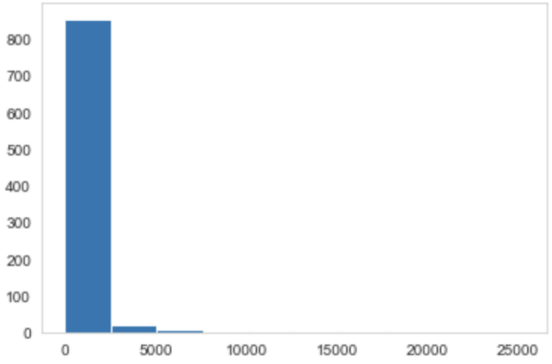


Figure (1) Histogram of Daily unique visitor

We visualized the variables in order to obtain clear information and examine the link between them. Traffic to a website is represented by the UV (unique visitor) indicator. Channels with high traffic are the preferred choice for businesses looking to advertise because they have a wider

audience and the opportunity to reach more people. Figure (1) represents the histogram of the Daily_UV, the x-axis represents the range of value of daily unique visitors and the y-axis represents how many observations fall into each range. The majority of daily_uv are found in the 0 to 2500 range, indicating that the majority of channels have medium to low traffic.

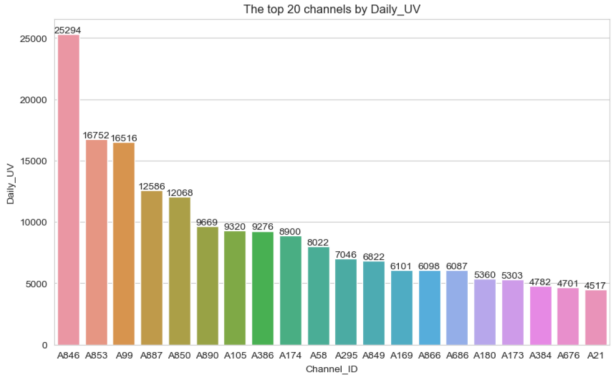


Figure (2) The Channel ID of top 20 daily unique visitors

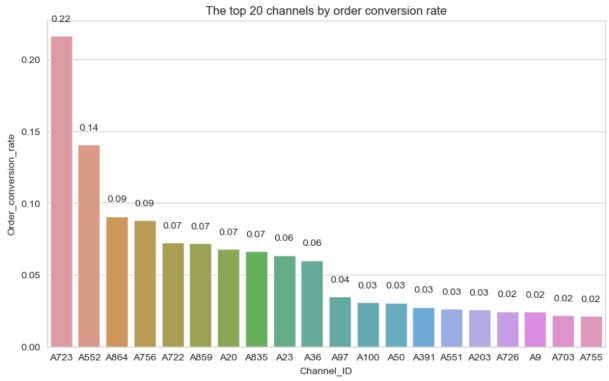


Figure (3) The Channel ID of top 20 order conversion rate

In Figure (2), we see that the top 5 Channels with over 10,000 daily unique visitors and the Channel A846 have 25294 daily unique visitors. The order conversion rate of the Channel A846 is 0.0019, which seems that the high traffic will not decide the high conversion rate. Now, we explore the order conversion rate and the relationship between it and the daily unique visitor.

In Figure (3), three levels clearly distinguish the top 20 channels by order conversion rate. Channels having a conversion rate of more than 10% are included in the first level. A552 and A723 fall in the first level, with A723 having a higher than 20% order conversion rate. Channels having a conversion rate of between 5% and 10% are included in the second level. In this range, there are a total of 8 channels. And Channels having a conversion rate less than 5% are included in the third level. In this range, there are ten channels altogether.

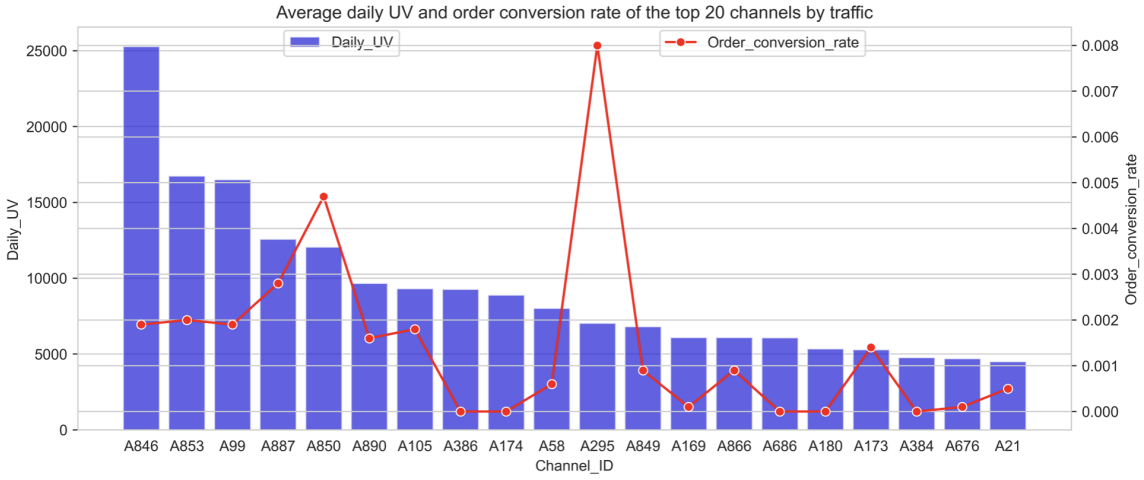


Figure (4) The Channel ID of top 20 daily unique visitor and their order conversion rate

In Figure (4), all of the top 20 channels by traffic have order conversion rates around 1%, showing that despite their enormous traffic, these channels are not extremely targeted and that the people they draw do not have a high level of readiness to pay. Next, we also want to know if a longer total advertising duration leads to better traffic results. In Figure (6), the Channel A723, with the highest conversion rate, has only been activated for 5 days. This indicates that there is almost no correlation between the two. But, we see that 65% channels in the top 20 order

conversion rate have been activated over 15 days. Does this also suggest that showing an advertisement to the public for a longer period of time will increase the likelihood that people will be drawn to it and make purchases?

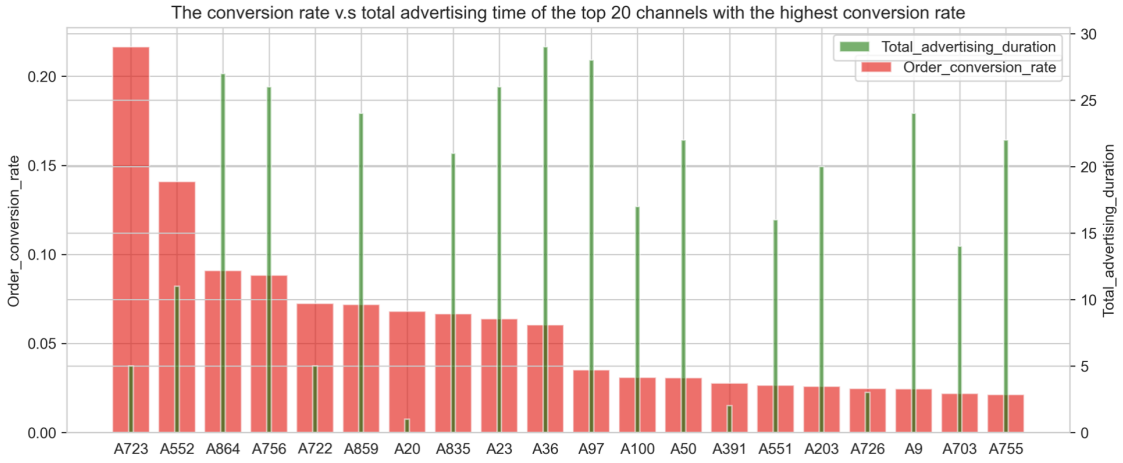


Figure (6) The Channel ID of the top 20 order conversion rate and their total ad duration

Let's look at the order conversion rate and daily unique visitors when the total advertising duration is 30 days (the longest time in this dataset). In Figure (7), the fact that the majority of them have a daily UV of less than 1000 and that just two of them have an order conversion rate of more than 1% suggests that there is still a ton of space for improvement in the advertising strategy for these channels due to the 30 day activation of these channels.

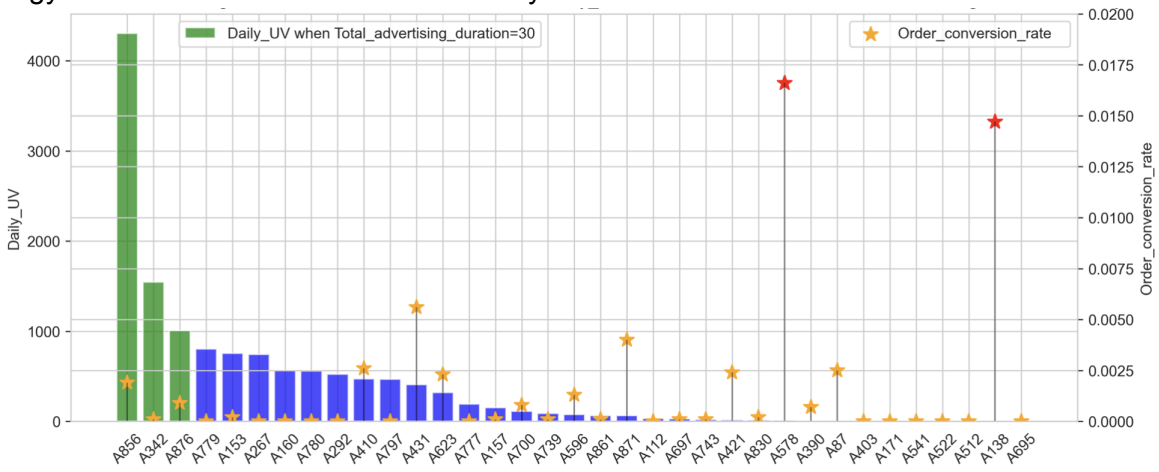


Figure (7) Channels has been activated for 30 days and their daily uv and order conversion rate

The channels that require attention are A856, which has a fair amount of traffic, as well as A578 and A138, which have fair amounts of channel conversion rates.

Now, let's look at the user behavior variables. Figure (7) shows that the top 20 channels have average registration rates that are higher than 1%, with channel A446 having the highest average registration rate at just under 4%. Despite having the highest average registration rate, channel A446's daily UV and order conversion rate show that neither the channel's ability to draw traffic nor its audience's willingness to pay is particularly strong. This is shown by the channel's low daily UV and order conversion rate. However, there is a fair amount of eagerness to sign up on the internet.

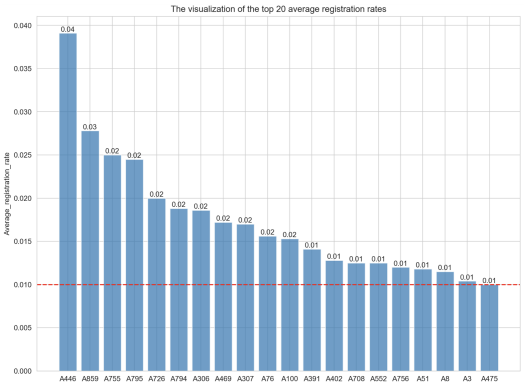


Figure (8) Channels with top 20 Avg_registration _rates

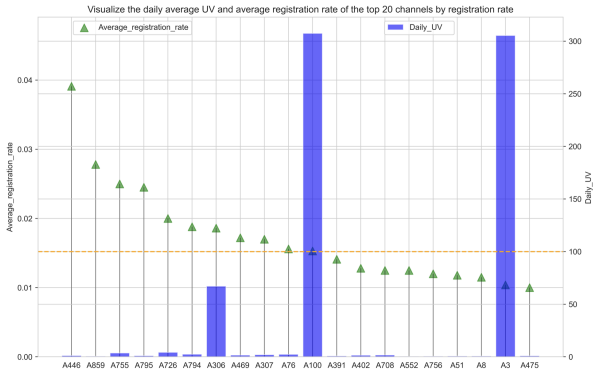


Figure (9) Channels with top 20 Avg_registration _rates and their daily uv

Figure (8) shows that the top 20 channels' average registration rate is greater than 1%. The highest-ranking channel, A446, has a registration rate that is generally close to 4%. And Figure (9) also shows there is no significant correlation between the average registration rate and the daily average UV, and the daily average UV of the top 20 channels with the highest average registration rate is generally not high. There are only two channels, A100 and A3, with a daily average UV exceeding 100.

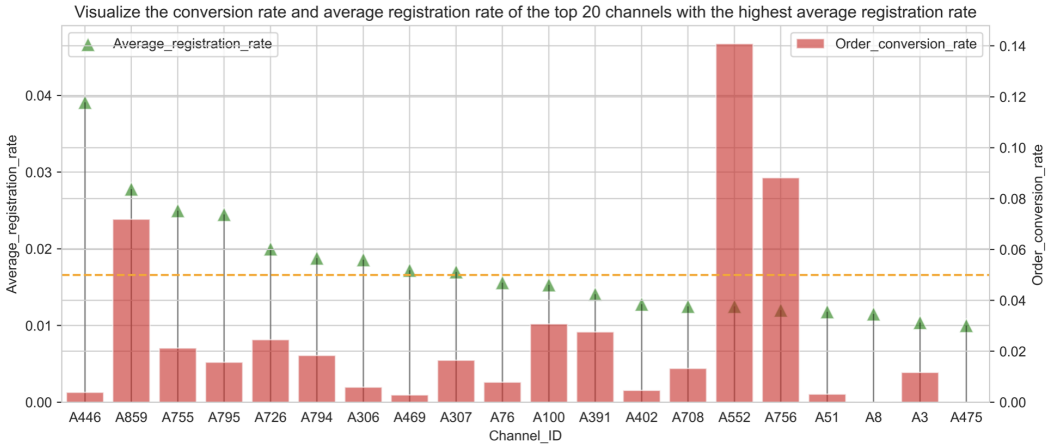


Figure (10) Channels with top 20 avg_registration_rate and their order conversion rate

Users who have registered on the website are more likely to make a purchase because there is a 31.65% correlation between the average registration rate and the order conversion rate. The order conversion rate is generally less than 5% among the top 20 channels with the highest average registration rate. Only 3 channels A859, A552, and A756 have an order conversion rate of more than 5%. With a daily UV of about 0.5, these three channels do not have a lot of traffic, and it is likely that they are low traffic platforms in vertical fields. Next, we calculate the correlation between the average session duration and order conversion rate, the correlation is 25.30% and the visualization is shown in Figure (11). The longer a person remains on a website, the more probable it is that they will make a purchase, so it seems sense that the average session time and order conversion rate are correlated. Additionally, of the top 20 channels with the longest average session length, channels A864, A20, A723, A756, and A859 have order conversion rates that are higher than 5%.

Now, let's think about what kind of advertisement can attract the people more. From the Figure(12), we can see that two sorts of media make up the majority of the ad material: jpeg and video. Additionally, only 4 channels are GIF-based. In Figure(12), Cpc is the primary one out of

a total of 4 cooperation techniques. A website owner or publisher is paid by an advertiser using the CPC (Cost-Per-Click) advertising approach each time a user clicks on their ad.

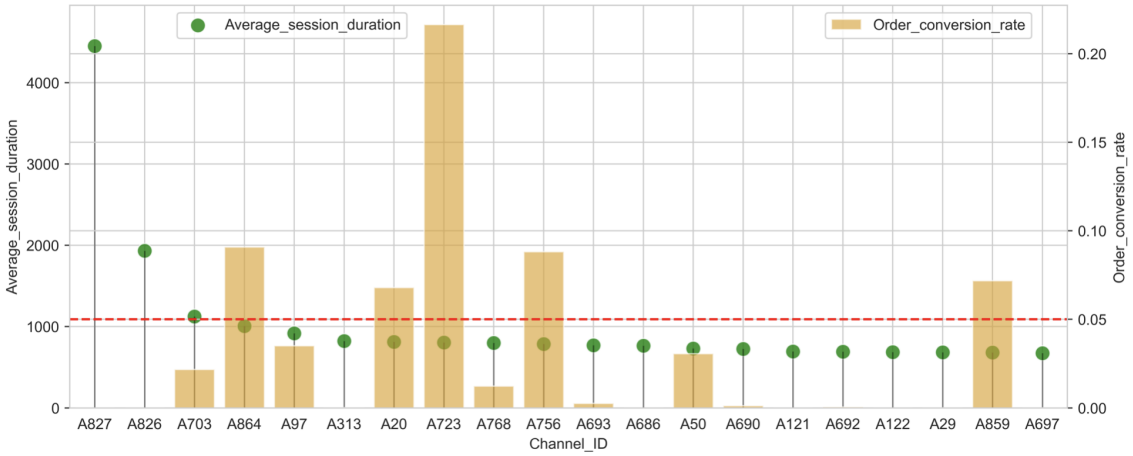
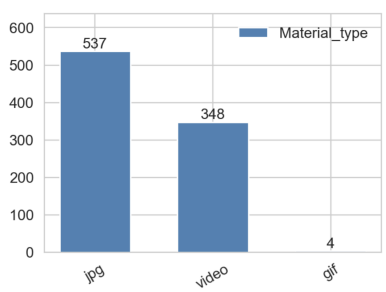
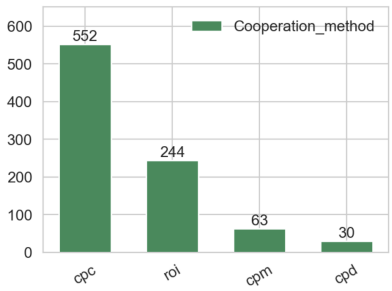
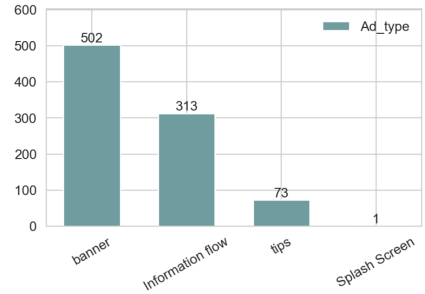


Figure (11) Channels with top 20 avg_session_duration and their order conversion rate

Return on Investment, or Roi, is the term used to describe how much an advertiser must pay the owner of a website. CPM stands for "Cost per Mille" or "Cost per Thousand Impressions". Regardless of whether the ad is clicked or results in a conversion, the advertiser agrees to pay a specific sum of money for each thousand times that it is shown. "Cost per Day" is referred to as "CPD." Regardless of the quantity of impressions, clicks, or conversions the ad generates, the advertiser commits to paying a set sum of money for each day that it is displayed.

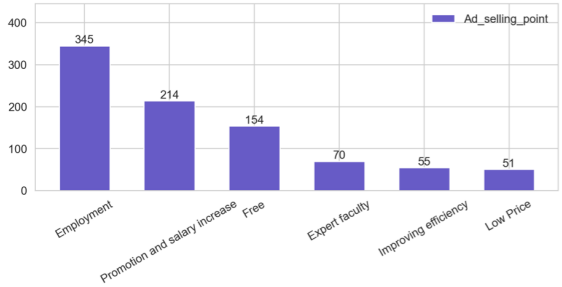


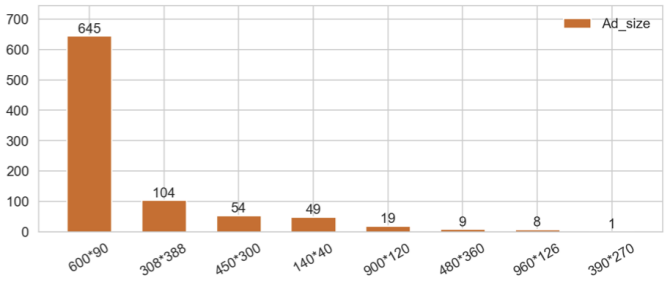




Figure(12) Histogram of Material type Figure(13) Histogram of Cooperation method Figure(14) Histogram of Ad type

In Figure (14), the two most common types of advertisements are banner and information flow. There are a total 6 kinds of selling point, and the top 3 advertisements are about employment, promotion, and free. In Figure (16), the majority of ads have a size of `600*90`.





Figure(15) Histogram of Ad_selling_point Figure(16) Histogram of Ad_size

III. Data Cleaning and Preprocessing

First, we used the info() function to check whether our data contains null values. We can see that the average session duration column contains 2 null values. To handle these missing values, we examined the corresponding records. Based on business judgment, we decided to fill the missing values with 0. The average registration rate and average search volume of these two rows are 0, and the visit depth is 1, indicating that the users attracted by these two channels basically click on the advertisement, take a look and leave, so it can be considered that the average stay time is 0. After filling in the missing values, we used the info function to double check and make sure the whole data again does not contain any null value.

Next, since there is a strong correlation particularly linear relationship between the visit depth and the average session duration. To avoid redundancy in subsequent modeling, we chose to keep only one of these variables. Considering the correlation coefficients, we found that the average session duration exhibits stronger correlations with other characteristics such as order conversion rate, daily UV, and average registration rate. Therefore, we decided to retain the average session duration. This led to a change in the data shape from (889, 13) to (889, 12). The data at this time consisted of discrete variables (object type) and continuous variables (float type).

To further process the data, we extracted all text-based data and started by handling the discrete variables. We used one-hot encoding to convert categorical variables, such as Meterial_type, Ad_type, Ad_size, etc., into a format appropriate for analysis or modeling. The get_dummies() method in Pandas makes it simple to one-hot encode categorical variables in a DataFrame. This resulted in the final feature matrix being 889 rows by 32 columns.

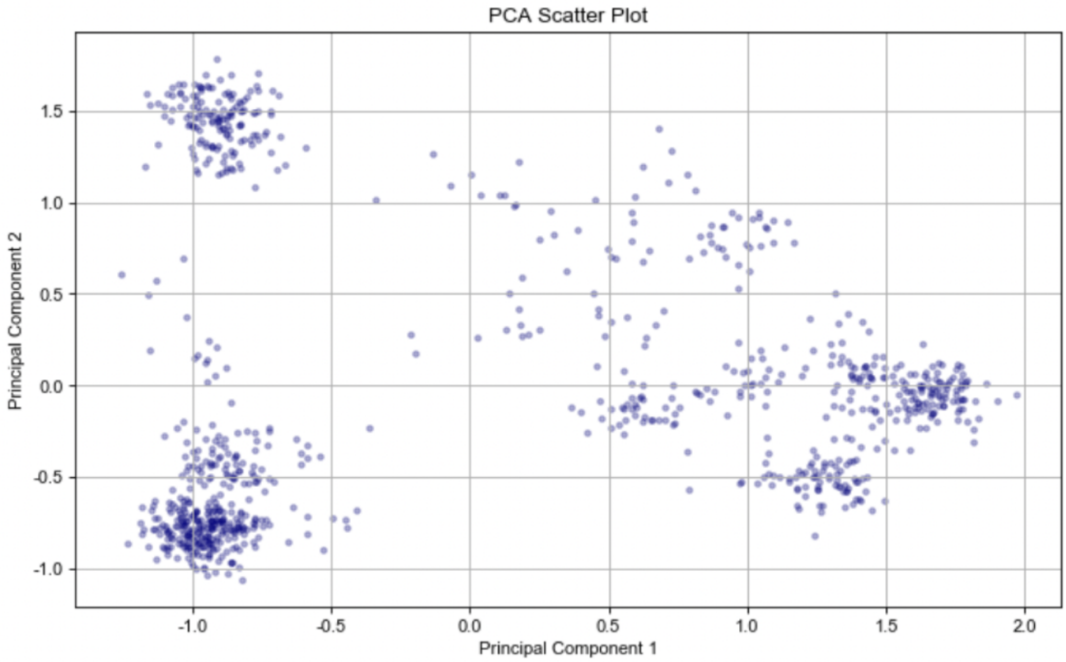
It is critical to handle the issue with various scales in the data since we intend to group the channels using the K-means clustering algorithm. When using K-means clustering to compare data points based on their proximity, non-uniform dimensions may have an impact on the clustering outcome. To normalize the continuous variables, we utilized MinMaxScaler. MinMaxScaler is a type of feature scaling in machine learning that scales the features to a fixed range of [0, 1]. It achieves this by subtracting the minimum value of the feature and dividing by the range of the feature. Standardization is another type of feature scaling that transforms the features to have zero mean and unit variance. It involves subtracting the mean of the feature and dividing by the standard deviation of the feature. But we chose to use MinMaxScaler in this case.

IV. Modeling

Cluster analysis is a technique that can classify data points or samples with similar characteristics in a large dataset into categories or clusters. It is particularly useful for data preprocessing when dealing with large sample sizes. Principal Component Analysis (PCA) is a popular technique for visualizing data following dimension reduction. With as much of the original variance preserved as possible, PCA attempts to convert the high-dimensional data from the original into a lower-dimensional space. In our case, we set the n_components parameter to 2, which means we retain the first two principal components. We did this because visualizing data in two dimensions is easier than visualizing data in higher dimensions, and the first two principal components often capture a significant amount of the variance in the data.

For the visualization, we have made adjustments to enhance the clarity of individual data points. We specifically reduced the marker size to 10 (s=10) and raised the transparency to 0.3

(alpha=0.3). These adjustments aid in improving the visual clarity of the clustering results by making it easier to distinguish between individual points in the scatter plot in Figure(17).



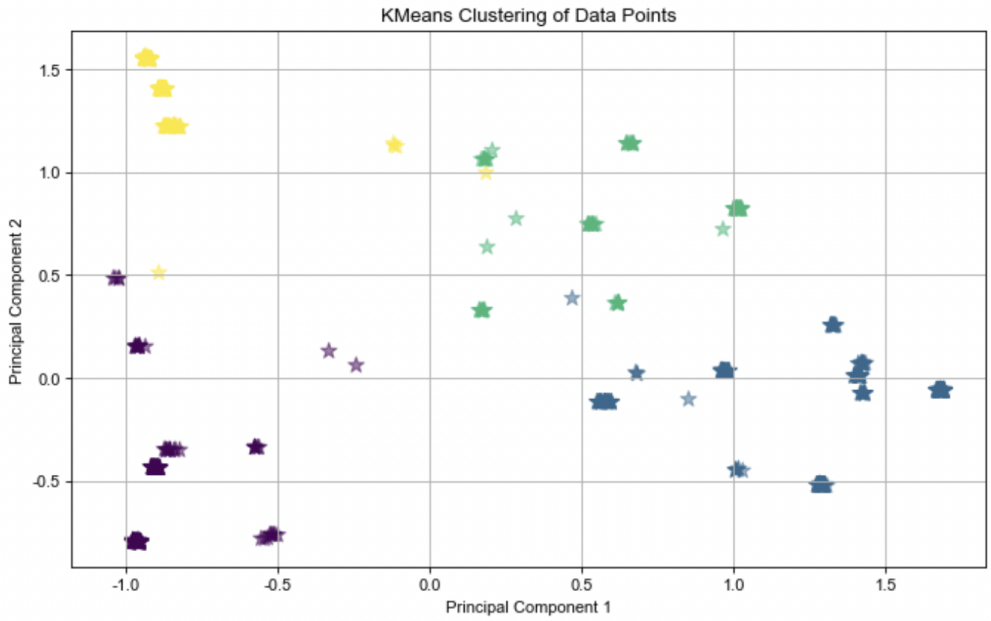
Figure(17) PCA Scatter Plot

To determine the optimal number of clusters for a K-means clustering model, we used the average silhouette score. The silhouette score, which ranged from -1 to 1, determined how similar an object was to its own cluster in comparison to other clusters. A score of 1 indicated a high similarity to its own cluster, a score of -1 showed a dissimilarity, and a score of 0 denoted a cluster boundary. A higher average silhouette score indicated better clustering results. In order to calculate the average silhouette score, we iterated across a range of cluster sizes from 2 to 7, building a K-means clustering model for each number of clusters. The K value with the highest average silhouette score was selected as the optimal number of clusters. We stored the silhouette scores for each K value in a DataFrame and printed them out. In our case, the best K value was 4, with an average silhouette score of 0.5016.

V. Results and Analysis

Finally, we created a scatter plot to visualize the data points after applying PCA and performing K-means clustering with K=4. This plot in figure(18) showed how the data points were grouped into different clusters based on their characteristics.

The results of each cluster group can be summarized by observing the mean value of continuous variables and the mode of discrete variables. We may combine the raw data with the clustering labels to examine the features of each cluster, and then compute the sample size and proportion for each cluster category. In addition, by choosing the data that correspond to each cluster index in the K-means model, we may compute the mean for numeric characteristics and the most prevalent value for non-numeric attributes. By concatenating these features, we can create a list of cluster-specific characteristics that represent the unique attributes of each cluster group; the final list is in the figure (19).



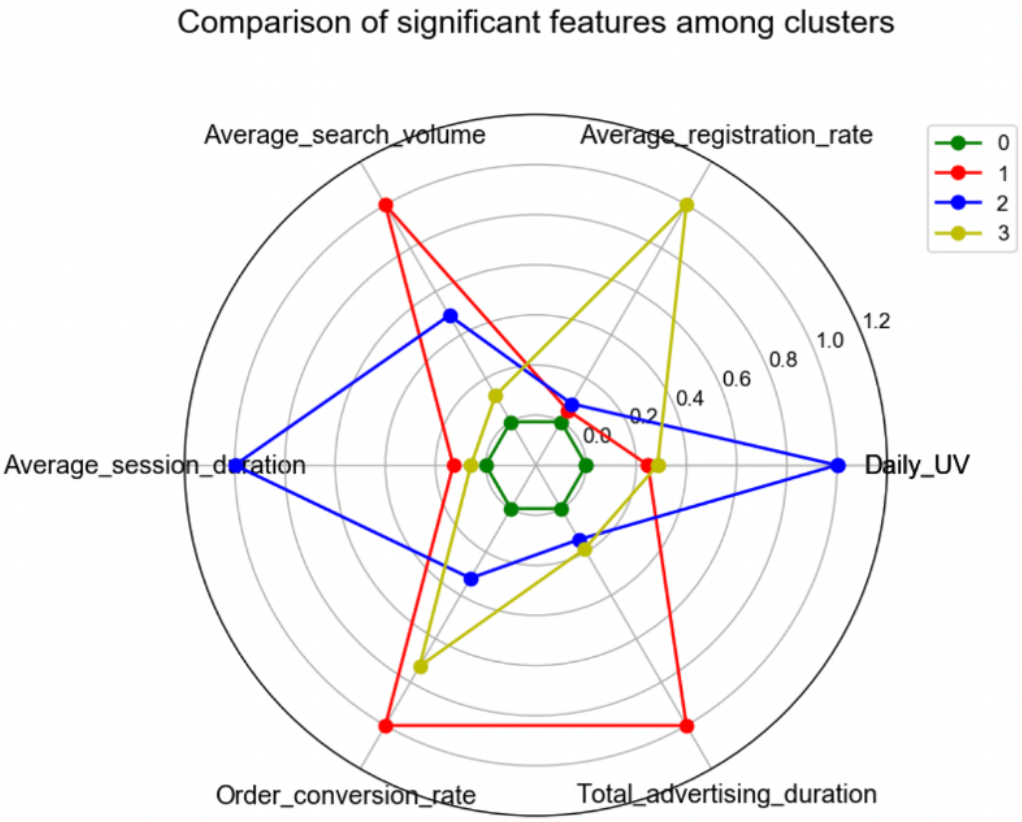
Figure(18) KMeans of Clustering of Data points

	0	1	2	3
counts	349	313	73	154
percentage	0.392576	0.352081	0.082115	0.173228
Daily_UV	300.205415	572.521054	1401.524521	613.835779
Average_registration_rate	0.001099	0.001182	0.001229	0.002649
Average_search_volume	0.01578	0.051364	0.033232	0.020118
Average_session_duration	237.134785	260.547796	418.084932	247.76539
Order_conversion_rate	0.001817	0.003868	0.002475	0.003309
Total_advertising_duration	15.34957	17.124601	15.60274	15.681818
Material_type	jpg	video	video	jpg
Ad_type	horizontal scroll	information flow	tips	banner
Cooperation_method	cpc	roi	cpm	cpc
Ad_size	600*90	600*90	450*300	308*388
Ad_selling_point	Employment	Promotion and Pay Raise	Promotion and Pay Raise	Free

Figure(19) Clustering Result

We created a polar coordinate map in figure(20) to visualize the clustering results better. The data was normalized using MinMaxScaler to ensure consistent scaling across different features.The map displays the significant features for each cluster in a radial format, allowing for easy comparison and identification of patterns among clusters.

It can be seen from the clustering results that all channels are divided into 4 categories, and the sample size of each category is 349, 313, 73, 154, respectively, with corresponding proportions of 39%, 35%, 8%, 17%. From the polar map, we can see that the first type of channel (index



Figure(20) Polar Coordinate Map

value is 0) is in a disadvantageous position in all aspects of indicators, indicating that this type of advertising media belongs to the situation of "high input, low output" effect quality, and as it accounts for 39%, it is one of the main channels of advertising media. The second type of channels (index value is 1) is the relatively high order conversion rate and average search volume, and the weakness is the low daily UV, indicating that this type of channel has a high conversion effect but insufficient traffic.The third type of channels (index value is 2) has particularly prominent average daily UV and average session duration, which is more in line with the role positioning of "attracting new" in advertising media. The fourth type of channels (index value: 3) has high user registration rate, and above-average transformation effect, indicating that the effect of this type of channels on user registration and transformation is relatively comprehensive.

VI. Discussion

Understanding the characteristics and strengths of each channel type allows for strategic decision-making in allocating resources and designing effective advertising campaigns. For the first type of channel, the performance is relatively average in all aspects, and trade-offs need to be considered in subsequent advertising. For the second type of channel, the main characteristic is high conversion rate but low traffic, making it more suitable for user conversion, especially for improving order conversion. For the third type of channel: the main feature is large traffic, meeting the basic demand of advertising "wide coverage". Therefore, it is suitable for large-scale advertising and traffic generation. For the fourth type of channel, both user registration and conversion have good performance, making it a media with relatively good

comprehensive effect. Therefore, this type of channel can be considered for advertising in various scenarios.

In conclusion, this project studied different types of advertisements and utilized the K-Means algorithm to classify advertising channels based on three months of data. The clustering analysis revealed four distinct types of channels with varying performance in different metrics. These findings can assist in making informed decisions regarding advertising strategies. Future improvements for this project include exploring alternative clustering algorithms and conducting in-depth analysis on each cluster. By leveraging these findings, businesses can optimize their advertising strategies and improve user acquisition and return on investment.