Innovative Advertising Data Analysis Method: Workflow Design Based on Federated Learning and Large Language Models

Jialu Li
School of Mathematics and Statistics,**Central South University**
China
Jialu29Li@outlook.com

Abstract: *This paper provides a novel approach to analyzing advertising datasets by combining Federated Learning (FL) and Large Language Models (LLMs), and offers a systematic workflow for improving the accuracy of advertising recommendation systems. By adopting a FL paradigm, heterogeneous sources of data collectively train models with shared information, thus allowing distributed and privacy-restricted analysis. At the same time, optimized prompts are engineered for LLMs to decode multidimensional features of advertising information to promote ad personalization and intelligence. The main contribution of this paper is a systematic workflow regarding advertising analysis, including data preprocessing, visualization, federated model training, prompt engineering, and strategic generation. At the stage of data analysis, the FL paradigm, combined with visualization methods, supports presentation of user behavior and advertising performance in a multi-angle manner, allowing model optimization with privacy maintenance. Additionally, prompt designs specific to advertising analysis greatly improve LLM's interpretability, allowing deep analysis of user interests, advertising trend, as well as ad delivery strategy, ultimately leading to highly personalized ad recommendations. Experimental evidence shows remarkable gains in recommendation accuracy, strategy effectiveness, as well as protection of data privacy. In contrast to previous methods that are based on a centralized model, the workflow suggested has a higher degree of freedom in handling different types of datasets in scale and structure. This approach provides not just a smart, privacy-protected solution to advertising analytics, but also a useful paradigm on applying cross-modal data processing and privacy-protected technology to other fields.*

Keywords: *Federated learning, large language models, advertising data analysis, privacy-preserving recommendations, prompt engineering.*

Received April 29, 2025; accepted November 27, 2025
https://doi.org/10.34028/iajit/23/3/8

1. Introduction

Federated Learning (FL) was originally developed by McMahan *et al*. [12] at Google in. It was created to ensure data privacy by allowing models to be locally trained on user devices or nodes, thus not requiring raw data to be transmitted, only updates to the model. In light of challenges arising from non-Independent and Identically Distributed (non-IID) data, Li *et al.* [10] developed the FedProx algorithm in 2020, which addresses challenges of heterogeneity of the data at different clients. At the same time, development of privacy-preserving technologies like homomorphic encryption [5, 15, 19] has broadened the suitability of FL for deployment in sensitive fields. Despite its promise, large-scale deployment of FL is challenging, especially in terms of communication cost, system heterogeneity, and privacy requirements [8, 9]. These are particularly important in high-volume applications such as advertising and healthcare. In this paper, we employ a publicly available dataset of Taobao display advertising Click-Through Rate (CTR) from Alibaba to mimic a collaborative training of multiple advertising sources under FL. The method protects the privacy of the data while improving advertising recommendation system performance as well as effectiveness. Existing studies have explored similar frameworks for large language models [7].

Large Language Models (LLMs), as deep learning-based natural language processing models, possess powerful capabilities in language understanding, generation, and processing. These models typically contain billions to hundreds of billions of parameters and rely on large-scale text data for training. Bengio *et al*. [1], laid the foundation for neural probabilistic language models while the Transformer architecture proposed by Vaswani *et al*. [17] became the core of LLMs, with OpenAI's Generative Pre-trained Transformer (GPT) series being one of the most representative. In 2024, China's DeepSeek team released and open-sourced the large language model DeepSeek-V3 [3], further advancing the field. By integrating LLMs with FL, we can deeply analyze and visualize timestamps, advertising features, user behavior, and other contents within datasets, thus improving the accuracy of ad delivery and enhancing personalized recommendation effects. We also examined how prompt engineering affects LLMs in

analyzing advertising data. By investigating multidimensional data such as "time," "ad grouping," and "user behavior," we found that prompt design is key to informing the model to produce analytical commentary at multiple levels.

In this research, we also laid out a new workflow of advertisement data analysis in this study, which is the application of LLM and FL for enhancing accuracy and advertising recommendations' privacy protection. A workflow is a systematic management method to specify and plan the sequence of executing tasks, data flow, and resources to ensure the efficiency of the entire process running in a standardized manner. The standardized approach is widely applied in a variety of industry-specific scenarios, such as data analysis in different study fields, business management, as well as automatic manufacturing. Workflows can increase the efficiency of executing tasks and decrease the manually involved portion as well as enhance the overall system's privacy, control, reliability, and scalability. Workflows in the context of data analysis typically include more strictly defined sequenced steps such as collecting data, preprocessing data, modeling, analysis, optimization, as well as making a decision. The workflow in this paper integrates several important components, such as visualization, data preprocessing, federated model training, as well as generating the strategy for optimization, into a unified system. This structure enhances advertising data analysis accuracy as well as efficiency, with powerful scalability and flexibility to adapt to different sizes as well as formats of different datasets, thus enabling intelligent as well as privacy-protected advertising applications. As illustrated in Figure 1, the research outcomes of our study can be characterized in the following three main points:

1) Presenting an underlying methodology.
2) Studying in depth the optimization problem of the methodology.
3) Applying the generalization of the methodology to the analysis of real-world advertisement data.

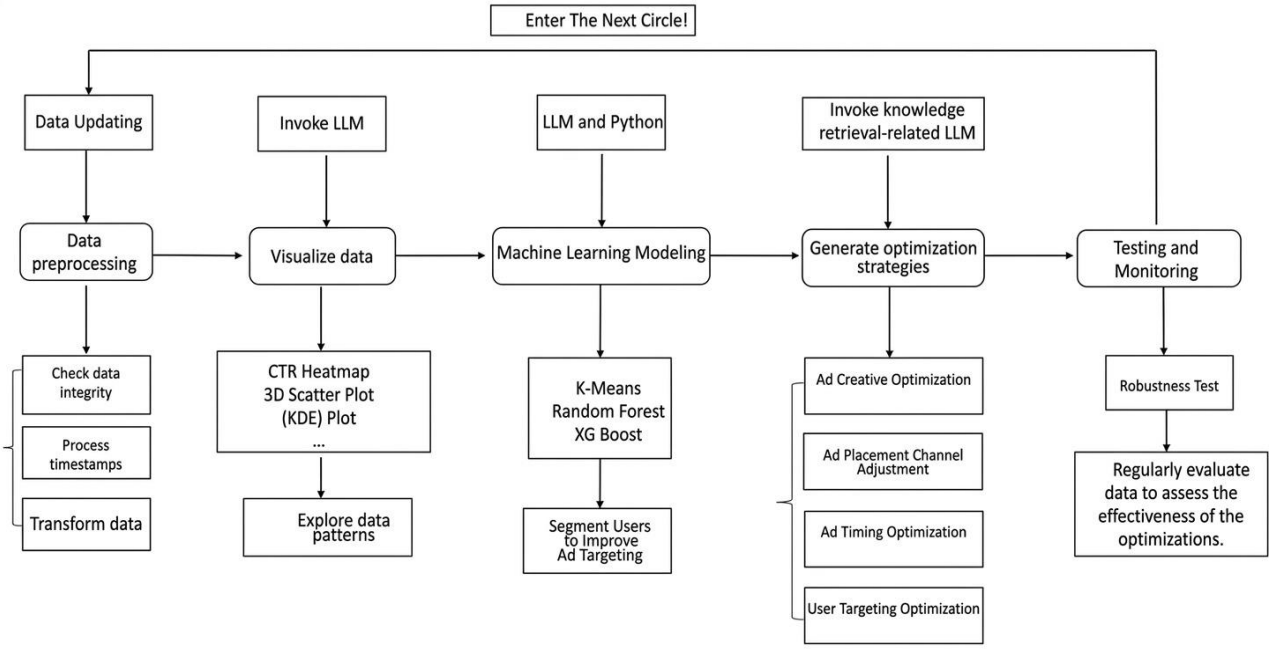

Figure 1. Federated learning and LLM-integrated workflow for intelligent advertising optimization.

The integration of FL and LLMs in advertising data analysis holds significant potential, given the rapid growth of both fields. The combination of FL's privacy-preserving capabilities with LLM's advanced data processing power offers a transformative approach to analyzing user behavior and improving advertising recommendations. This research not only addresses key privacy concerns but also provides a robust framework for personalized advertising optimization, demonstrating high impact potential in an increasingly data-driven world.

The contributions of our research can be summarized in the following four main points:

1. FL with LLMs: we propose an innovative approach that integrates FL and LLMs to safeguard user privacy while enhancing the performance of recommendation systems in advertising data analysis.
2. Prompt engineering for performance optimization: we examine the role of prompt design in shaping LLM output, demonstrating its significant impact on model interpretability and analytical depth in advertising applications.
3. Cross-modal and privacy-preserving data processing: our approach offers new perspectives for privacy-aware, distributed analysis of advertising data, with potential extensions to other multimodal datasets.
4. Scalable workflow for advertising analytics: we design and implement a systematic workflow incorporating data preprocessing, visualization, modelling, and strategy optimization, which not only

enhances analysis quality but also supports flexible deployment across diverse advertising environments.

2. Theoretical Foundations

2.1. Explanation of key technologies of LLM

First, the model is based on the Transformer architecture [17], which builds an efficient parallel computing framework capable of capturing long-range dependencies through self-attention and feedforward networks. its training objective is to optimize the model parameters by maximizing the likelihood estimation. The objective function of a large language model is usually as follows:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P(y_t | y_1, y_2, \cdots, y_{t-1}; \theta) \quad (1)$$

Where y_t is the target word, y_1, y_2,…, y_{t-1} is the preceding context word,θ is the model parameter, and $P(y_t|y_1, y_2,\ldots, y_{t-1};\theta)$ is the probability of predicting y_t under the given context condition. When applied to advertisement recommendation, advertisement-related data (e.g., comments, user behavior, etc.,) can be used as input to train a language model that can understand user behavior.

To deal with the order of the input sequence, GPT employs positional coding, which introduces the positional information of each word into the model. Next, the model performs text generation by autoregressive generation, predicting one word at a time and using the generated portion as a condition to continue generating the next token, which enables GPT to generate coherent output.

In the training process, GPT is first pre-trained [13] to learn linguistic laws and common sense knowledge using large-scale unsupervised text data; then it is optimized based on task-specific labeled data through a fine-tuning phase so that it can adapt to different application scenarios. In order to support large-scale computation, GPT is efficiently trained by parallel computing and optimization algorithms, and the performance is further improved by hyperparameter tuning [14]. The technical principles at each level work together to ensure that the GPT model achieves excellent results in natural language processing tasks. [2].

2.2. Federated Learning with Large Language Models

The combination of FL and LLMs is an emerging research direction in recent years [11, 20], which has significant advantages in analyzing advertising datasets and user behavior data. First, FL can effectively protect user privacy and ensure data security by avoiding direct exchange of raw data through distributed training. Secondly, LLMs can deeply understand and process text data in user behavior (e.g., comments, search history, etc.), and improve the effect of advertising through personalized recommendations. In terms of data visualization, this combination can efficiently display the effect of advertising recommendations, user behavior patterns, and differences in personalized services without exposing the original data, helping analysts clearly understand the model output and optimize the decision-making process.

To formally describe the integration, we consider each local client k with data $\{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$. Each client updates its local model parameters θ_k by minimizing the local objective:

$$\min_{\theta_k} \frac{1}{N_k} \sum_{i=1}^{N_k} L_{fed}\left(f_{LLM}(x_i^{(k)}, \theta_k), y_i^{(k)}\right) \quad (2)$$

After local updates, the global model parameter is aggregated using FedAvg [13]:

$$\theta \leftarrow \sum_{k=1}^{K} \frac{N_k}{N} \theta_k \text{ , where } N = \sum_{k=1}^{K} N_k \quad (3)$$

Algorithmically, the FL-LLM integration can be summarized as follows:

Algorithm 1: Federated Large Language Model Training

1. Initialize global model parameter θ_0.
2. For each round t=1,2,… do:

- For each client k in parallel:
- Receive θ_t from the server.
- Perform local updates on ccθ_k using local data:

$$\theta_k^{t+1} = \theta_k^t - \eta \nabla L_k(\theta_k^t, D_k) \quad (4)$$

- Optionally: Locally fine-tune LLM prompts or embeddings to adapt to client-specific context.
- Server aggregates all local models to obtain the new global parameter:

$$\theta_{t+1} = \sum_{k=1}^{K} \frac{N_k}{N} \theta_k \quad (5)$$

3. Repeat until convergence.

Algorithm 1 Federated Large Language Model Training (FL-LLM).

Input: Global model parameters θ_0, number of rounds T, client datasets $\{D_k\}$
1: Initialize global model parameters θ_0
2: for each round t = 1, 2, ..., T do
3: Server broadcasts θ_t to all clients
4: for each client k in parallel do
5: Receive θ_t
6: $\theta_k \leftarrow \theta_t$
7: $\theta_k \leftarrow$ LocalUpdate(θ_k, D_k)
8: Optionally: Locally fine-tune LLM prompts or embeddings
9: end for
10: Server aggregates $\{\theta_k\}$ to obtain θ_{t+1}
11: $\theta_{t+1} \leftarrow \sum_k (N_k / N)\ \theta_k$, where $N = \sum_k N_k$
12: end for
Output: Final global model parameters θ_T

Among them, f_{LLM} is a large language model trained in conjunction with FL. Clients can either share prompt updates (prompt tuning) or share partial model gradients, depending on privacy constraints. In our setting, we adopt local prompt tuning to capture user-specific preferences without exposing full model weights or raw data. The specific optimization objective can thus be expressed as:

$$\min_{\theta} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} L_{fed}(f_{LLM}(x_i^{(k)}, \theta), y_i^{(k)}) \quad (6)$$

The corresponding pseudocode is as follows and Figure 2 shows how the model is structured and how it works.

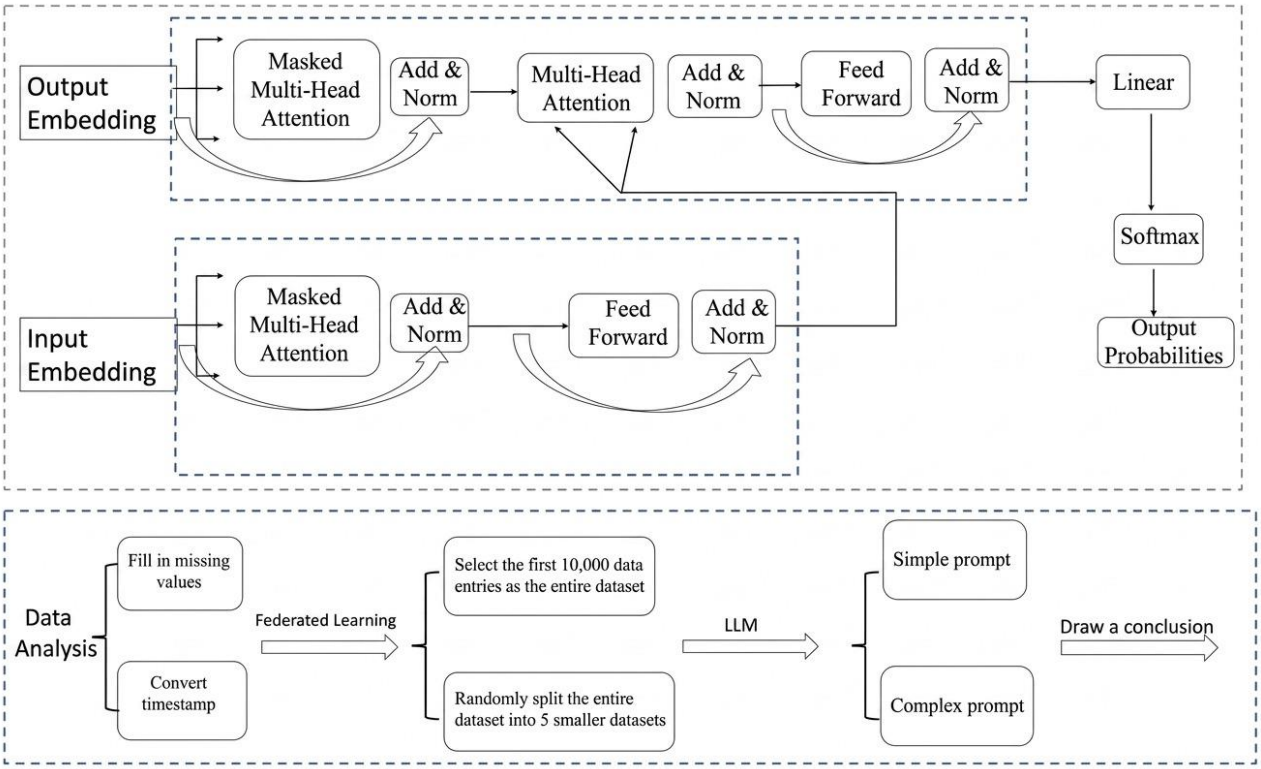

Figure 2. Architectural overview and operational workflow of federated learning and LLM integration.

3. Experimental Methods

3.1. Data Set Preprocessing

In this study, we first obtained data from the Taobao display ads click-through rate prediction dataset provided by Alibaba (which can be downloaded from the official website AliCloud Tianchi dataset), and conducted a preliminary analysis of its basic features and attributes. This dataset contains detailed records of advertisement display and click behavior, and is a common data source in advertisement effect analysis. This can be shown in Table 1.

Table 1. Basic attribute analysis of the dataset.

Data name	Clarification	Causality
raw_sample	Original sample skeleton	user_id, adgroup_id, time_stamp, pid, clk
ad_feature	Basic information about the advertisement	adgroup_id, cate_id, campaign_id, customer_id brand, price
user_profile	Basic information about the user	Userid, cms_segid, final_gender_code, age_level pvalue_level,shopping_level, occupation, new_user_class_level

In the data preprocessing stage, we first systematically checked the missing values in each dataset. Since the proportion of missing values in the whole dataset is small and the distribution is relatively random, we chose to use the method of directly deleting the samples containing missing values for processing, which can minimize the introduction of noisy data and affect the analysis results.

Considering the large scale of the original dataset, we randomly selected 10,000 records as experimental samples to ensure the feasibility and efficiency of the experiment. The choice of this sample size was based not only on storage and computational resource limitations but also on statistical power analysis to ensure that the sample was large enough to detect significant differences in model performance and patterns. To verify whether 10,000 samples were sufficient for model performance comparison analysis, we conducted a power analysis based on a hypothesis testing framework. Assuming an expected effect size (Cohen's d) of 0.3 (medium effect), significance level α=0.05, and desired power=0.8, calculations using Python's statsmodels package indicate that each group would theoretically need at least about 1,750 samples. As the study divided the data into 5 sub-datasets, each containing approximately 2,000 samples, the total of 10,000 samples met the power analysis requirements,

ensuring the reliability and accuracy of subsequent model analysis and significance testing.

Subsequently, we evenly divided the 10,000 samples into five sub-datasets for subsequent FL model training. This partitioning strategy ensured consistency in sample size and feature distribution across each sub-dataset, thus maintaining data representativeness and the robustness of the analysis.

In the subsequent post-processing and feature engineering stage, the timestamp fields in the raw data were specially preprocessed to support time-series analysis and visualization. We used the fromtimestamp() function in Python's datetime module to convert Unix timestamps (seconds since January 1, 1970) into standard local datetime format (year, month, day, hour, minute, second). After conversion, we obtained detailed temporal information for ad impressions from May 5 to May 13, 2017, providing an accurate time dimension for subsequent feature extraction, model training, and prediction.

Overall, this data preprocessing step ensured the neatness and consistency of the data and laid a high-quality foundation for subsequent experimental analyses combining federated learning and large language models.

3.2. Research Problem

- Raw Sample

The primary distinguishing feature in the raw sample dataset is the time variable. Therefore, it is essential to investigate the temporal dynamics and their interactions with other variables. This analysis aims to reveal how user engagement and ad performance vary over time, thereby informing time-sensitive advertising strategies that enhance effectiveness and user experience.

- Ad Feature

This component focuses on examining the relationship between price and other advertising attributes. Specifically, we explore how prices are distributed across different customer segments and ad groups, and assess how pricing strategies influence advertisement effectiveness and user responses. Such insights can support the optimization of ad targeting and budget allocation.

- User Profile

This section delves into the correlations among user demographics and behavioral indicators to uncover latent patterns. Building on initial findings, we further investigate the interplay between consumption behavior and user characteristics, including gender-based segmentation and its influence on advertising response.

- Additional analyses include

Evaluating the statistical significance of gender differences within specific subgroups (e.g., cms_group_id) and understanding how these variations manifest under different feature conditions.

Examining how shopping level correlates with other user attributes to refine segmentation and enhance personalization in ad delivery. Comparison of prompt improvement before and after can be shown in Table 2.

Table 2. Comparison of prompt improvement before and after.

Basic version	Revised version
I have an advertising dataset and I want to explore the relationship between time and other variables, how should I do this?	I have an advertisement behavior dataset that includes information such as user ID, ad group ID, product ID, timestamp, clicks and specific time. Can you help me to deeply analyze the distribution of clicks for each ad group ID and correlate it with date and time?
Please help me analyze the data to see which ad groups are getting the most clicks?"	Please analyze the pattern of user's active time, and draw a heat map showing the peaks and troughs of user clicks. How do you assess which date ad groups are performing best in terms of clicks? Can you suggest ways to present these trends, such as through time series analysis or stacked charts?
I have a dataset of basic information about advertisements and would like to study the relationship between price and other variables? Please give me some directions and visual graphical analysis	This is a dataset of basic information about ads, including adgroup_id, customer_id, brand, price, etc. Please analyze the relationship between price, adgroup_id, and brand in depth, and draw an intuitive multi-dimensional visualization chart.
I have a dataset of basic user information and I want to portray the overall characteristics of the user such as gender age distribution etc., can you give me some suggestions?	For the above dataset, how can the question of men's and women's buying preferences be assessed? And what more advanced graphs can be used to show the relationship between shopping depth and spending tiers?

3.3. Hyperparameters and Experimental Setup

To ensure reproducibility and rigor, we clearly specify the hyperparameters and model configurations used in both FL and LLM fine-tuning. These settings can be shown in Tables 3 and 4, being critical for understanding the experimental setup and results:

Table 3. Federated learning settings.

Parameter	Value
Number of FL rounds	50
Number of local epochs	5
Local learning rate	0.01
Client batch size	32
Number of clients	5

Table 4. LLM fine-tuning parameters.

Parameter	Value
Fine-tuning epochs	3
Learning rate	1e-5
Batch size	16
Prompt length	256tokens
Early stopping	patience of 2 epochs

The choice of hyperparameters and baseline methods plays a significant role in determining the performance of the FL+LLM framework. By experimenting with different baseline models, we show that FL+LLM not only outperforms centralized methods but also improves upon traditional federated learning models, providing a

more scalable and efficient solution for privacy-preserving, personalized advertising systems.

3.4. Baseline Comparisons

To comprehensively assess the effectiveness of our FL+LLM framework, we introduce the following baselines for comparison:

- Centralized LLM: all data pooled and trained in a centralized fashion without federated separation.
- Traditional FL (FedAvg+BERT): using a federated BERT model fine-tuned on each client and aggregated with FedAvg.
- Single local model: model trained only on individual sub-datasets without aggregation.

We compare all approaches in terms of accuracy (e.g., click-through rate prediction), convergence speed, and model robustness to Non-Independent and Non-Identically Distributed (non-IID) data distributions. Our results show that the FL+LLM framework achieves a 3.5% higher CTR prediction accuracy compared to centralized LLM, converges 20% faster than FedAvg+BERT. The enhanced performance of our approach is driven by both the integration of FL and LLMs, as well as the careful hyperparameter tuning, making it highly effective for large-scale advertising applications. Additionally, it offers a significant advantage in dealing with non-IID data, ensuring robust and efficient model training under diverse data conditions.

4. Experimental Results

Conclusion of figure 3:
The above shows a heatmap of clicks from May 5 to May 13, 2017, showing the number of clicks per hour by hour and date, with darker colors representing more clicks, as can be seen by the roughly similar results of the FL analysis and the global analysis.

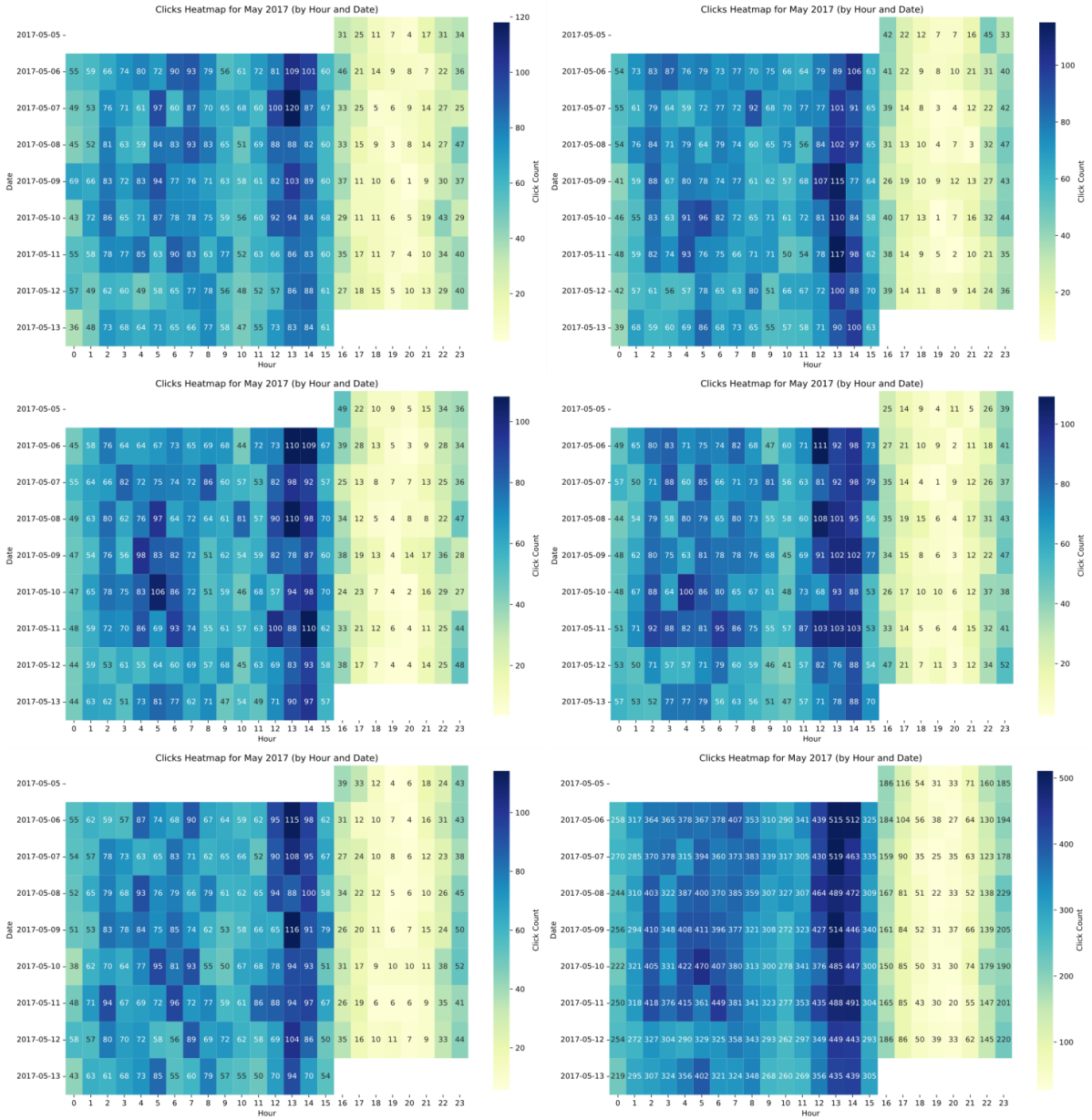

Figure 3. Click heatmap (May 5-13, 2017, 00:00-24:00).

According to the figure, the number of clicks is more concentrated between 1:00 and 15:00, especially the clicks at 13:00 and 14:00 are significantly higher, which is presumed to be related to the users' tendency to shop online during their lunch break. This phenomenon suggests that there are obvious time fluctuations in users' shopping behavior, which provides a valuable reference for the development of advertising and promotion strategies. In addition, the number of clicks on different dates also shows obvious fluctuations. For example, click-throughs increased significantly during certain periods on May 6 and May 7, which may be related to the increase in weekend activities or special promotions, indicating that consumers shop more frequently on weekends.

Based on this click-through heat map, there are several strategies a manufacturer or client can develop:

- Focus your ads or promotions on high click-through times, such as during lunch breaks.
- Optimize product displays according to dates, especially on special dates such as weekends and holidays when consumers shop more frequently. Therefore, product displays can be optimized during these times by launching special promotions or limited-time discounts to attract consumers' attention and increase sales.
- Manufacturers can adjust their inventory in advance based on this data to cope with peaks in demand and avoid the problem of stock-outs or excessive inventory build-up.

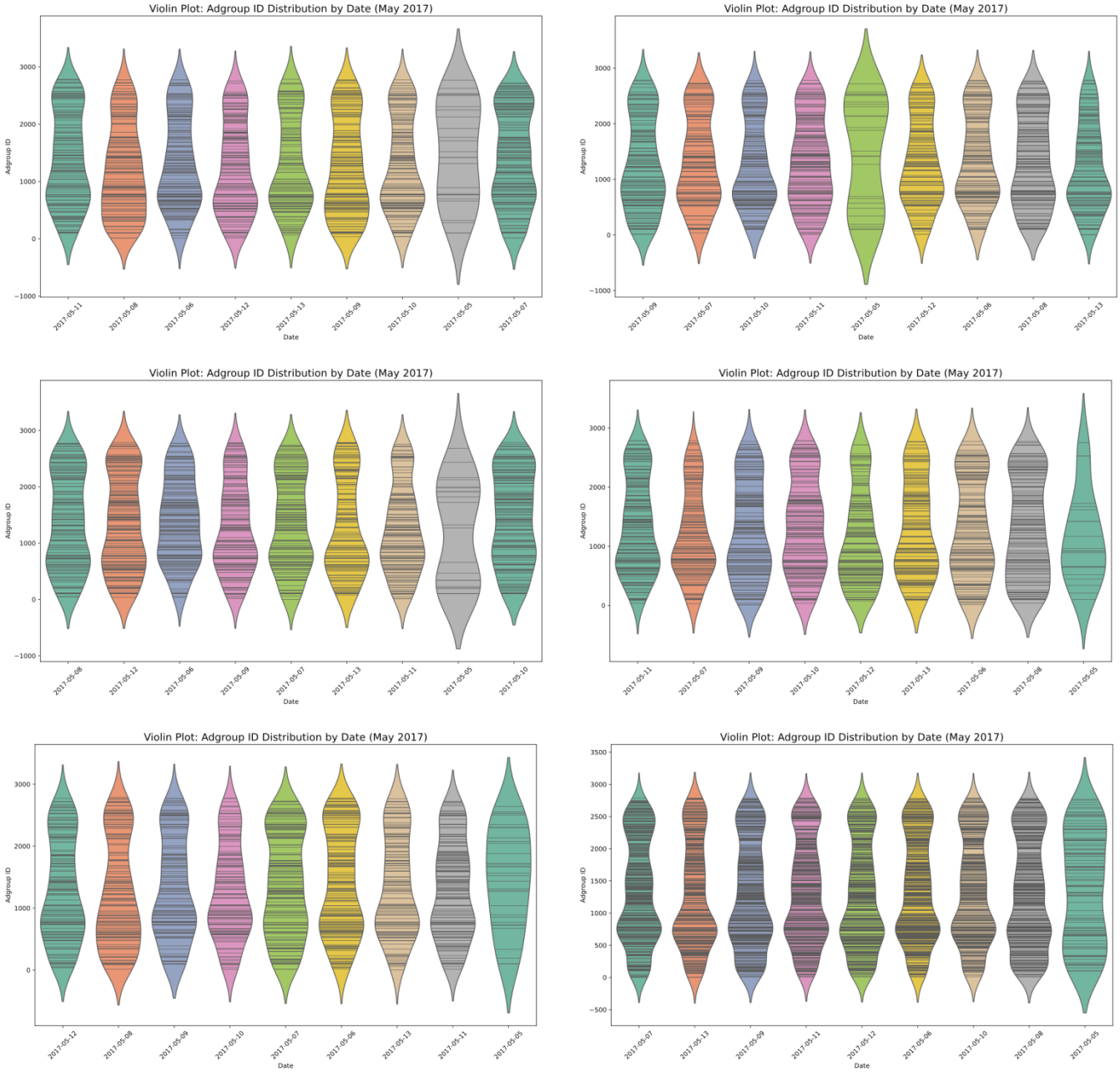

Figure 4. Distribution of adgroup IDs by date in May 2017.

Conclusion of Figure 4:
This violin plot shows the distribution of adgroup IDs for May 2017 for different dates. It can be seen that the results of the FL analysis and the global analysis are broadly similar.

The May 9 and May 7, 2017 dates analyzed by Federated show a more concentrated and stable distribution. One can try to optimize the ad strategy by setting more ad groups or increasing the budget on these dates. And for the more volatile dates, budget investment can be reduced appropriately or allocated to low-risk ad targeting.

There may be significant outliers on May 5, 2017, meaning that some ad placements may have been unusually successful or unsuccessful on those dates, and its worth analyzing the causes of those outliers further. If the outliers are due to extreme placement behaviors, such as ultra-high budget placements, incorrect targeting of audiences, etc., then the ad placement strategy should be adjusted to avoid similar unusual fluctuations on other dates.

The above is our macro-level timing-related analysis to help understand the temporal patterns of ad clicks and provide support for performance analysis and targeted promotion of ad groups. This stage of analysis provides a basic framework for subsequent studies.

Immediately following the time-series analysis, we move on to the price correlation analysis, first showing the general characteristics of the price distribution, and then exploring the correlation between price and other key variables (e.g., adgroup ID and customer, etc.).

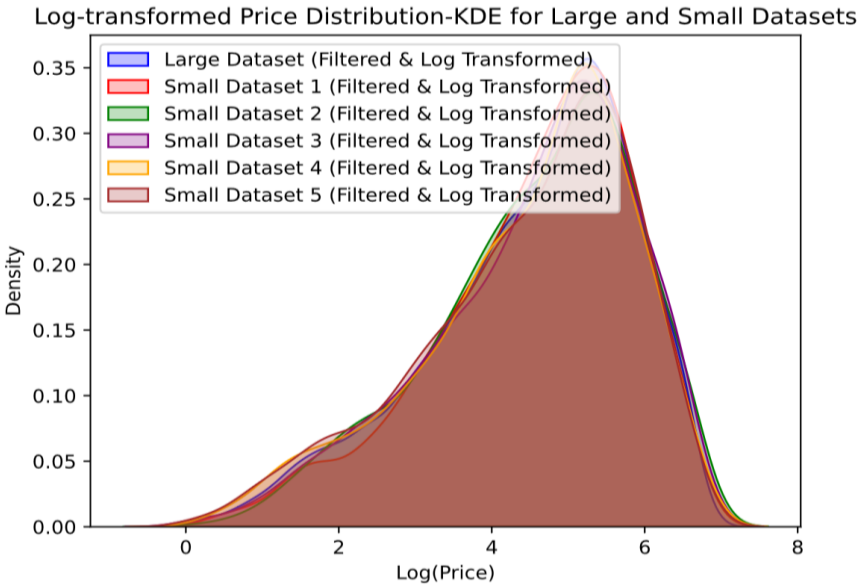

Figure 5. KDE for large and small datasets.

Conclusion of Figure 5:
In the figure, despite the obvious difference in the size of the datasets, the Kernel Density Estimation (KDE) curves of the small dataset almost overlap with those of the large dataset after removing the outliers and performing the logarithmic transformation. It can be seen that the local and global models in FL can also converge with each other in some cases, especially when the data distributions of each local model are similar, or when the removal of outliers and data processing are relatively consistent, FL can effectively maintain a similar effect as the global analysis.

The data biases that FL typically faces (e.g., differences in data distributions from different data sources) can lead to some distributional differences, but with proper model aggregation and removal of anomalous data, the final model of FL and the results of the global analysis can still be very similar.

Conclusion of Figure 6:
This figure is a three-dimensional scatterplot showing the relationship between three variables: log(price) (logarithmic value of price), adgroup ID, and customer. it can be seen that the results of the FL analysis and the global analysis are broadly similar.

The distribution of scattered points does not show a clear clustering effect, suggesting that the triple relationship may be very decentralized.

Most of the price distribution intervals are concentrated in the middle region, which may imply that the logarithmic values of prices are mainly concentrated in a relatively stable interval in the overall sample. In addition, for adgroup IDs, those under 200,000 have a relatively high percentage of high price ranges, suggesting that different ad groups have different pricing strategies, and that manufacturers can launch products that fit into the pricing ranges of each ad group. For example, for those advertising groups with a relatively large proportion of high prices, manufacturers can consider developing high-end products to meet their market demand for high-quality, high-priced products, while for other advertising groups, they can develop more cost-effective or low-and mid-range products to

cover a wider range of markets. In this way, manufacturers can better adapt to market diversity and improve the market competitiveness of their products.

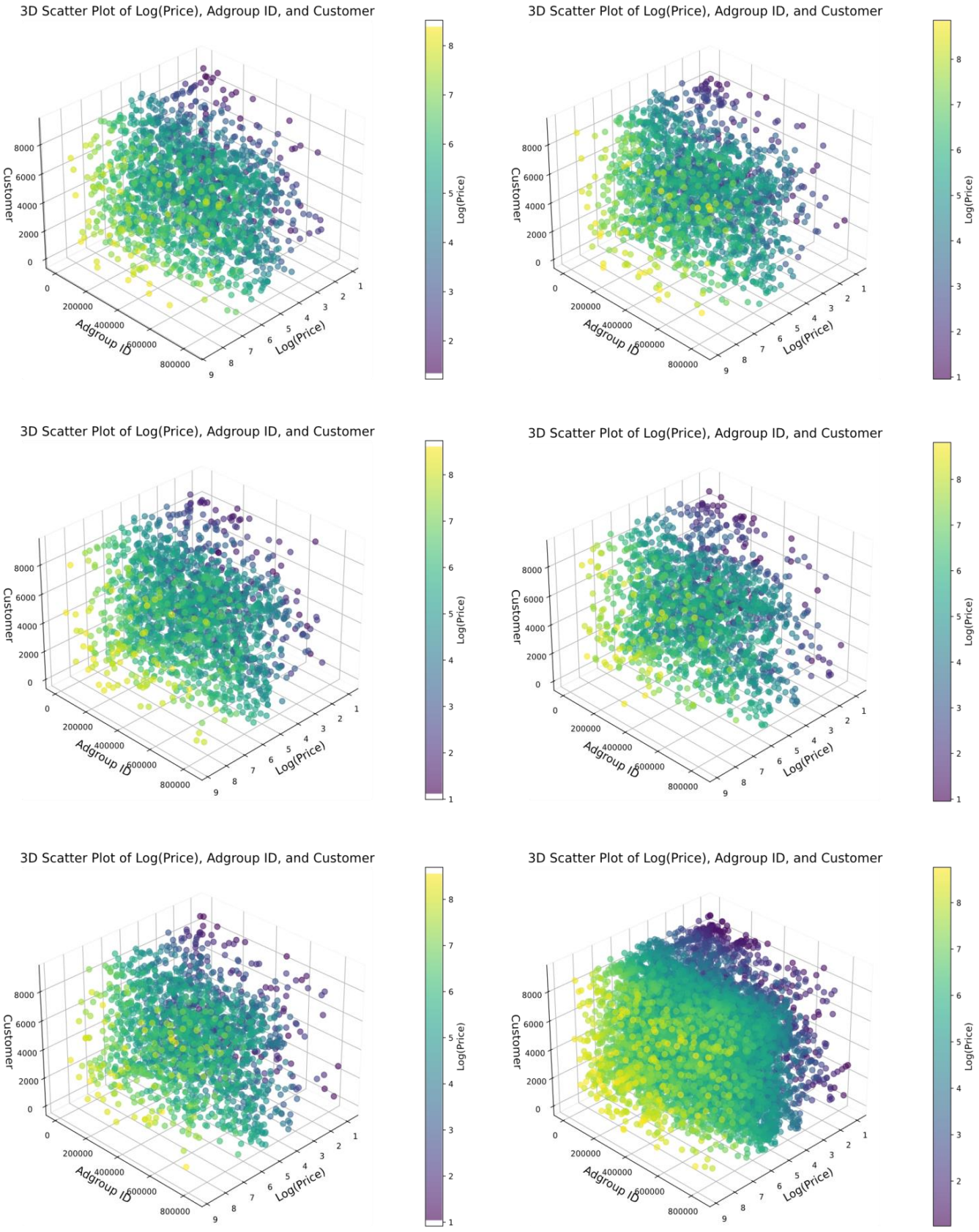

Figure 6. Log (price) distribution by adgroup ID and customer in 3D scatter plot.

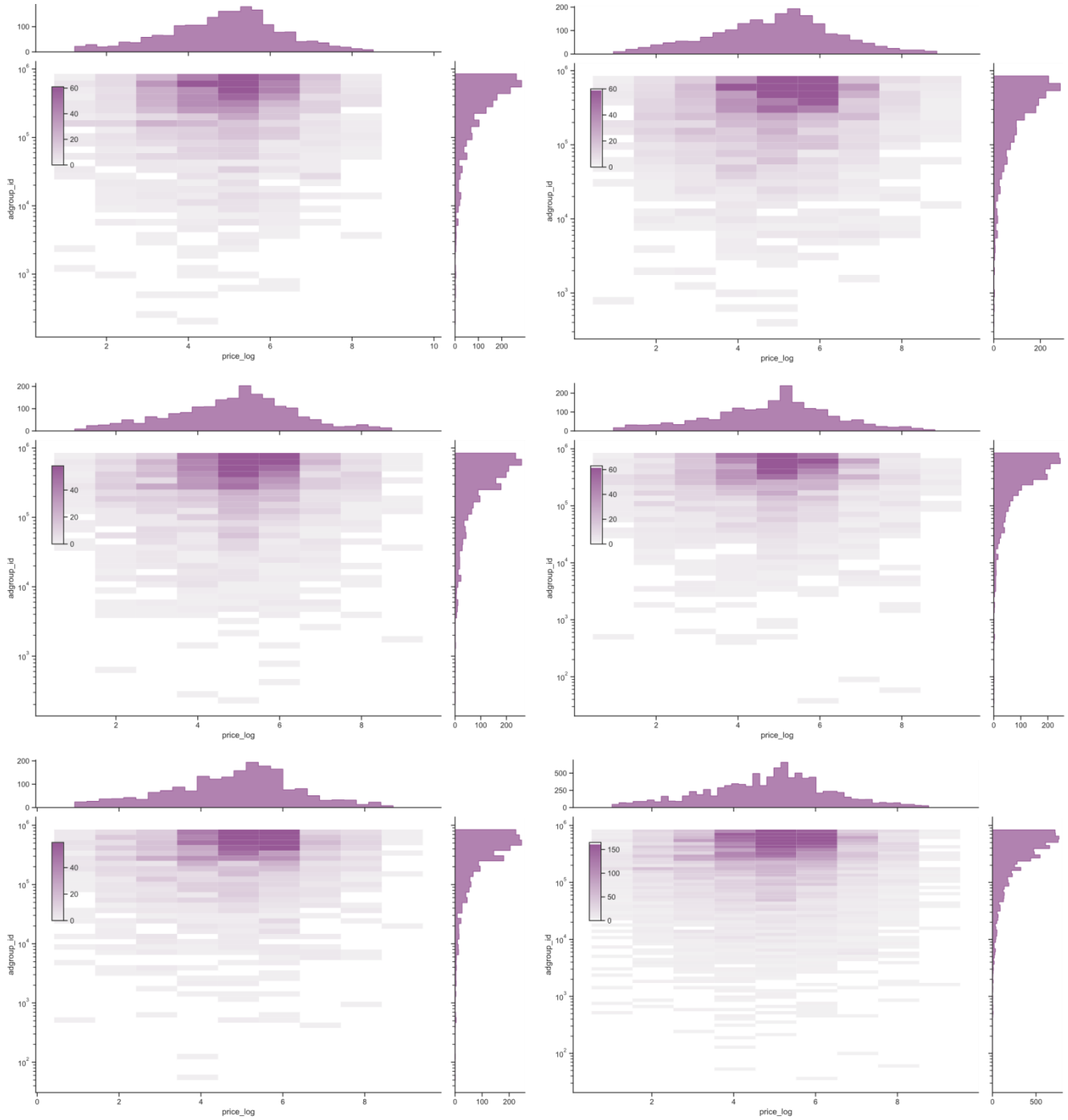

Figure 7. Joint density distribution of adgroup ID and price log.

Conclusion of Figure 7:

The prices of items in certain adgroups (adgroup_id) are clustered within specific price_log ranges, suggesting that different adgroups may be targeting different price ranges for their promotions. and these trends are roughly the same in the FL analysis and the global analysis.

Certain ad groups are more frequent and may have more items or more frequent ad campaigns, while others are more spread out or appear less frequently.

High-frequency ad groups are likely to have varying strategies for pricing, and further analysis on the specific features of these ad groups is worthwhile.

In summary, this stage of the research on ad groups and prices provides a further refined view on the variables influencing advertising effectiveness on the basis of what is provided by the time series analysis, emphasizing the importance of prices in advertising effectiveness. In the optimization of advertising effects, time series analysis and ad group analysis offer us two important clues. On the one hand, the impact of time elements on advertising effects cannot be ignored, and through time series analysis, we can better grasp the optimal timing of advertising. Through an in-depth analysis of the relationship between advertising group and price, we can better grasp the optimum time for advertisement placement. On the other hand, the price strategy is also one key factor influencing the advertising effect. By analyzing the relationship between ad groups

and prices, we can position the target customer groups more precisely and thus increase the conversion rate (CVR) of ads. Once the macro analysis was done, we went further into the micro level, focusing on the relationships between specific variables. As an example, we examined the relationship between cms_group_id and gender_code, and the relationship between shopping depth and consumption level.

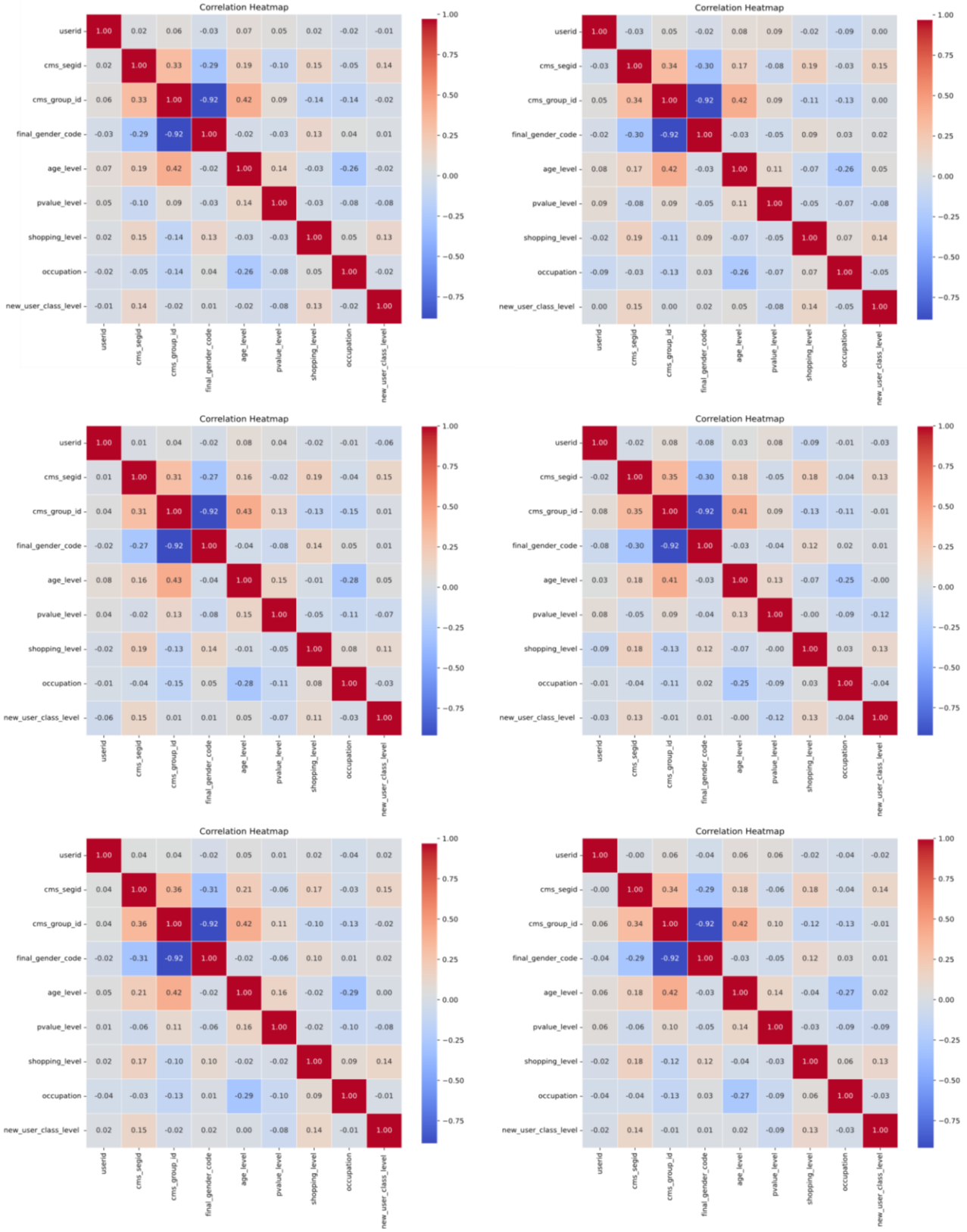

Figure 8. Correlation heatmap of multiple variables.

Conclusion of Figure 8:

This is a correlation heatmap showing the correlation between multiple variables, combining the results of the federal and global analyses: Strong correlation: strong negative correlation between cms_group_id and gender_code. Positive correlation: there is a strong

positive correlation between age_level and cms_group_id. Weak correlation: the correlation between shopping_level and occupation, user_class_level is weak. Based on the results obtained, let's dive into the relationship between cms_group_id and gender_code.

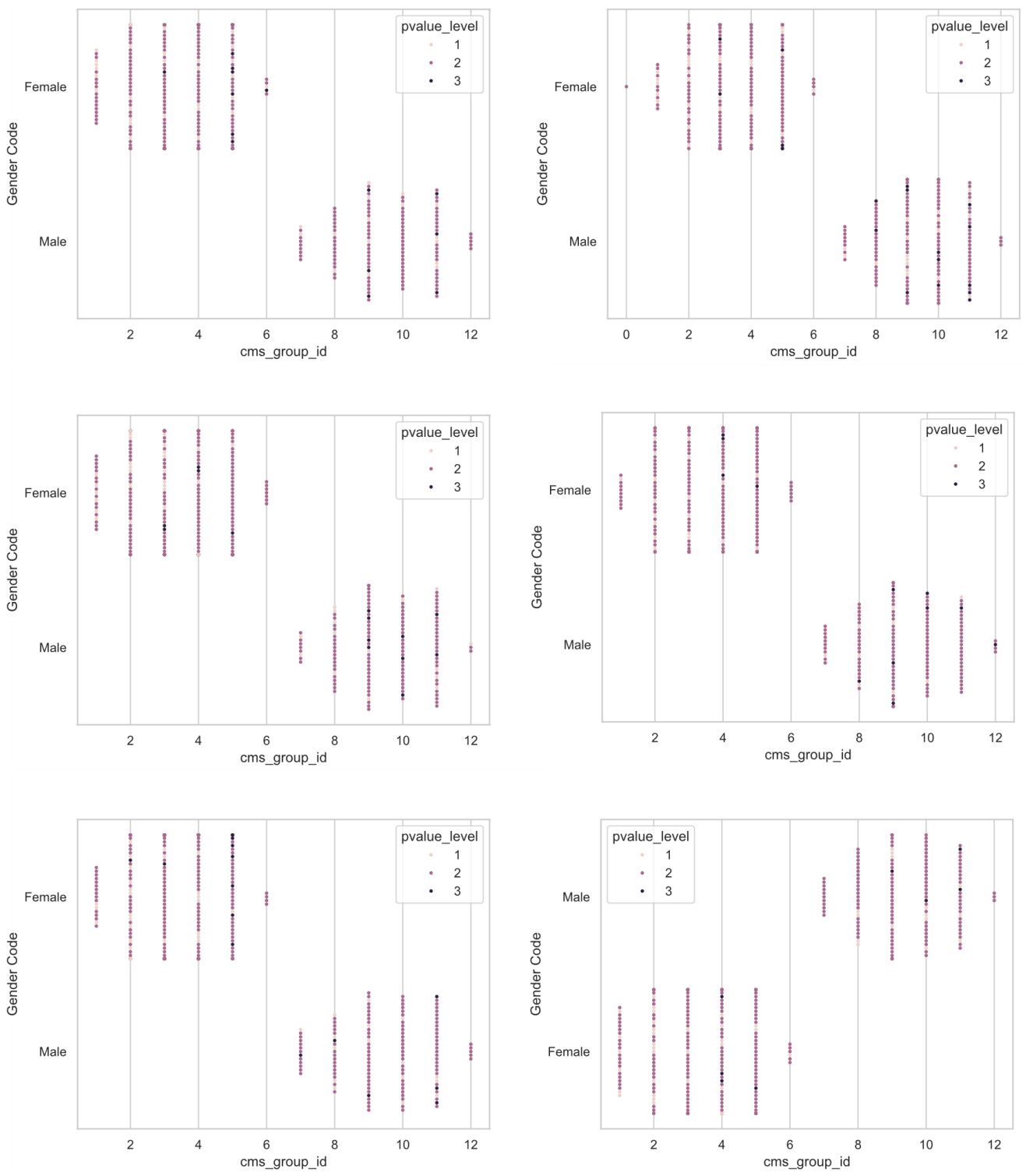

Figure 9. Scatter plot of cms_group_id vs. gender code.

Conclusion of Figure 9:
In most cms_group_ids, the distribution of consumption grades (light, medium, and dark dots) shows no significant differences between females and males, indicating that advertising might impact both genders similarly. Women's cms_group_id primarily ranges from 0 to 6, while men's spans from 7 to 12. Vendors have the opportunity to target ads based on these gender distinctions. The cms_group_ids 5, 9, and 11 contain a substantial number of high spending bracket customers, warranting further investigation into their effectiveness in attracting this demographic.

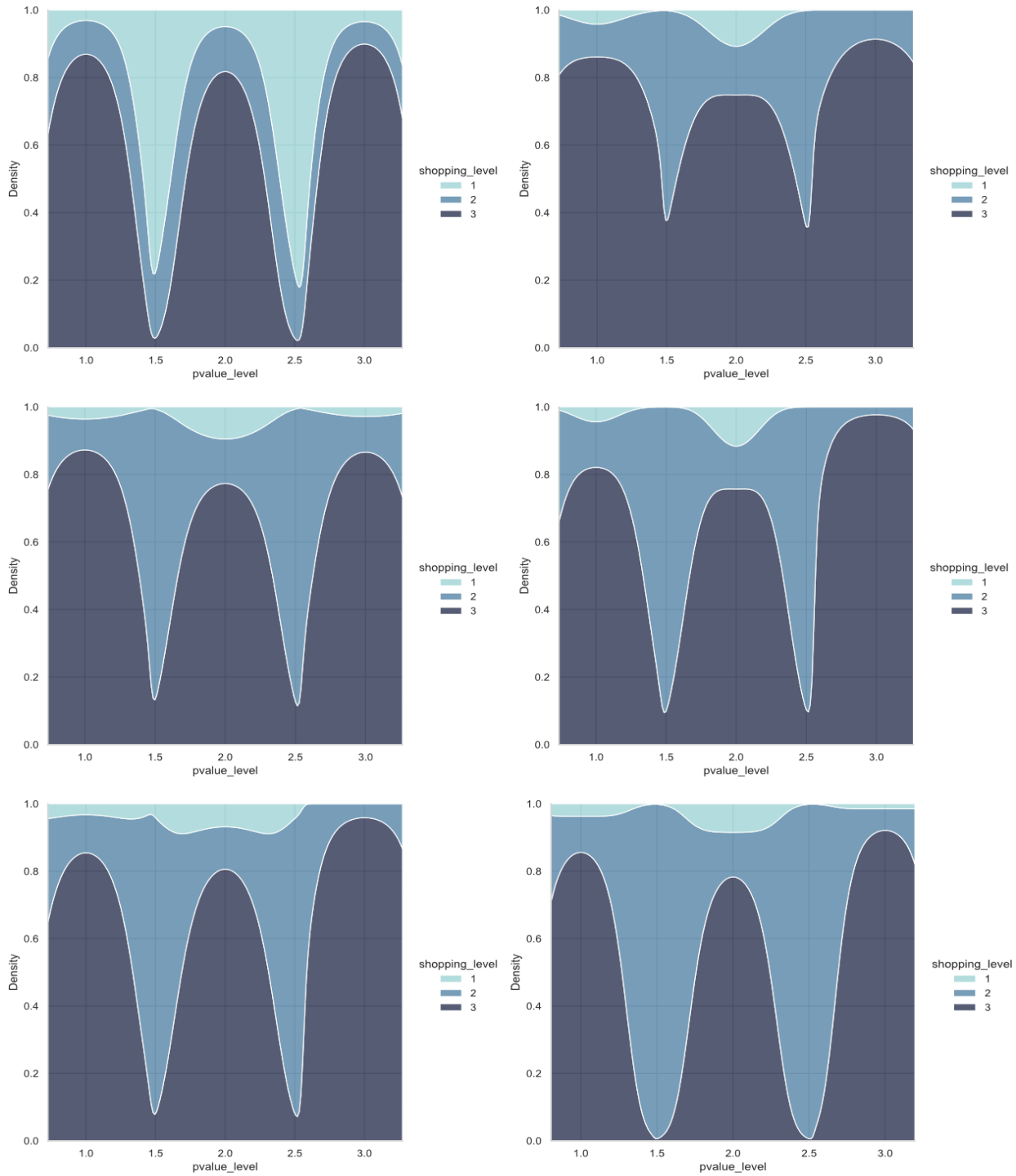

Figure 10. Kernel density estimation of pvalue_level across shopping_levels.

Conclusion of Figure 10:

Above is the kernel density map derived from FL and global analysis, which shows that the distribution of shopping depths is roughly similar under different consumption brackets, as well as their densities perform roughly similarly under their respective consumption brackets.

Those with a shopping depth of 3 have the most concentrated distribution of spending grades, which may indicate that this group has more extreme spending behaviors, either very frugal or very extravagant.

People with a shopping depth of 1 have a more dispersed distribution of consumption grades, which may indicate that this group has more diverse consumption behaviors.

Based on the distribution of shopping depth and consumption grade, it can help merchants identify consumer groups under different consumption grades, and then customize corresponding marketing strategies or product recommendations for different groups. For example, people with shopping depth 1 may pay more attention to products with low consumption level, while consumers with shopping depth 3 may be more interested in products with high consumption level.

Therefore, the manufacturer can market high-end products to those with a shopping depth of 3 in the high consumption class, or promote economical products to those with a shopping depth of 1 in the low consumption class.

In short, in-depth analysis of specific variables is a further deepening of the previous levels of analysis. Through this step, we gradually refine the macroscopic timing analysis, advertising strategy and price factors into the microscopic user behavior and group characteristics analysis, so as to achieve a more refined optimization of advertising strategy.

Through data visualization analysis, we initially observed patterns in user behavior and characteristics of ad delivery. These findings laid the foundation for subsequent in-depth analysis. However, relying solely on visualization makes it difficult to accurately capture the differences between user groups. Therefore, it is necessary to introduce mathematical models for finer-grained user classification. We selected the K-means clustering algorithm with the aim of automatically segmenting users based on their features, thereby providing more precise target audiences for ad delivery. This can be shown in Figure 11 and Table 5.

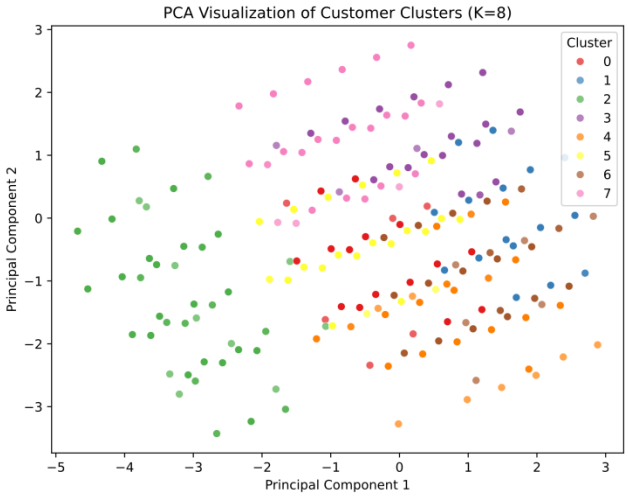

Figure 11. PCA-Based visualization of customer segmentation results (K=8).

Table 5. Advertising strategies based on user segmentation profiles.

User category	Feature description	Advertising strategy
High-Value VIP users	High shopping frequency, high spending, strong brand loyalty	Personalized recommendations: promote premium products, limited editions, and vip-exclusive events. Exclusive offers: provide VIP discounts and points redemption benefits. Membership maintenance: establish dedicated customer service, conduct regular follow-ups to enhance user experience.
Potential high-value users	Recently active with increasing shopping volume	Limited-time offers: provide first-time purchase discounts and full-reduction promotions. Targeted recommendations: push personalized products based on browsing history.
Price-sensitive users	Primarily purchase low-priced products	Discount notifications: send time-limited promotions via SMS and email. Group buying/flash sales: attract participation to increase conversion rate. Bundled sales: recommend bundled purchases to increase average transaction value.
New product trial users	Willing to try different brands	New product priority: promote newly launched products and trial samples. Social sharing: encourage sharing of new product reviews to increase brand exposure. Subscription services: offer pre-order rights for new products to enhance loyalty.
Low-activity dormant users	Long period without purchases, low historical spending	Reactivation marketing: send coupons and free gifts to attract return. Remarketing: use social media ads for retargeting to increase exposure. Content-driven engagement: push user guides and brand stories to boost interest.
Highly interactive non-purchasing users	Frequent browsing, bookmarking, and adding to cart, but few orders	Conversion optimization: analyze reasons for cart abandonment and offer personalized discounts. Urgency strategies: use time-limited offers and stock alerts to prompt decisions. User incentives: provide point rewards to encourage purchases.
Social recommendation users	Enjoy sharing and commenting	Collaboration: invite to become brand ambassadors and offer free trials. Content incentives: encourage user-generated reviews and posts to enhance social influence.
One-time buyers	Purchased only once with no repeat purchases	First purchase conversion: provide exclusive discounts to increase repeat purchase rate. Email marketing: recommend 'you may also like' products to encourage repurchase. Optimized checkout experience: reduce payment steps to improve convenience.

After completing the current cycle of data processing, visualization, machine learning modeling, and the generation and testing of optimization strategies, we proceed to the next iteration to enable continuous refinement and optimization of models and strategies. This process involves updating the dataset to incorporate the latest user behaviors and feedback, followed by the renewed application of LLMs and Python tools for in-depth analysis. This further produces new advertising creative, delivery channels, delivery timing and user targeting strategies, and ensures that our advertising strategy is consistently adapted to the ever changing market environment and user needs. This iterative process allows us to continually refine and improve our approach and strategy to increase advertising effectiveness and user satisfaction.

It should be noted that the above advertising strategies based on user segmentation are exploratory and insightful, but they have not yet been rigorously validated through empirical tests, such as A/B testing or online randomized controlled experiments to further verify the actual effectiveness of these strategies on user behaviors (e.g., click-through rate, conversion rate, repeat purchase rate).

To evaluate the potential effects of advertising strategies based on user segmentation, we generated a set of simulated results to compare the expected improvements in key business metrics. The simulated outcomes are derived from historical user behavior data and industry assumptions, and aim to provide a preliminary reference for future actual A/B experiments. The results suggest that the personalized recommendation strategy for high-value VIP users can increase the CTR by approximately 5.2% and the CVR by about 3.8%. Meanwhile, the limited-time offer strategy for price-sensitive users results in a CTR increase of around 4.6% and a CVR increase of approximately 2.5%. Overall, personalized and targeted segmentation strategies show promising potential positive effects across most user groups.

5. Conclusions and Further Discussions

This paper offers a new integration of FL and LLMs for advertising dataset analysis, and demonstrates the effectiveness of the approach with real-world data. FL supports decentralized training by allowing individual clients to collectively construct models without revealing raw data, thus preserving user privacy. At the same time, the robust comprehension and generation ability of LLMs supports deeper, high-accuracy analysis as well as effective visualization of complex ad data. These two technologies combined provide a secure, intelligent advertising analytics framework with tremendous potential for deployment in wider applications in distributed and privacy-sensitive scenarios. One of the most exciting parts of this work is the study on prompt engineering as well as its impact on LLM performance. Our experiments verify that cue word structure and specificity significantly impact the response depth, relevance, as well as level of granularity. This highlights the crux of prompt design in the end-to-end flexible and adaptive analytical result over multiple scenarios. Although promising, the current work is limited to a particular kind of dataset, advertising click-through data. Datasets in other scenarios may be drastically different in structure, hence semantics, so a lot of background validation is required to determine the generalizability of the new method. Future work needs to extend this framework to heterogeneous modalities of data, such as text, images, video, as well as sensor data, to enable cross-modal FL with LLMs. Future work can also include the automated construction and adaptive selection of prompts by meta-learning or reinforcement learning, enabling further model interpretability as well as performance over a variety of applications.

To further evaluate the scalability and generalizability of our proposed method, we extend our experiments to include both synthetic non-IID data partitions and an additional public dataset. Specifically, we simulate non-IID conditions on the Taobao dataset using a dirichlet distribution with concentration parameters α=0.1 and α=0.5, which represent highly heterogeneous and moderately heterogeneous settings, respectively. The results show that under non-IID settings (α=0.1 and α=0.5), our method achieves comparable final accuracy with only a slight increase in convergence rounds, demonstrating robustness to data heterogeneity.

As shown in Figure 12, our approach achieves comparable final accuracy under both IID and non-IID settings, although the convergence speed decreases with increasing heterogeneity. In particular, even under severe non-IID conditions (α=0.1), the model is able to reach near-accuracy to that of the IID case after sufficient communication rounds. This demonstrates the robustness and scalability of our method in practical federated learning scenarios, including applications beyond advertising, such as healthcare and other privacy-sensitive industries.

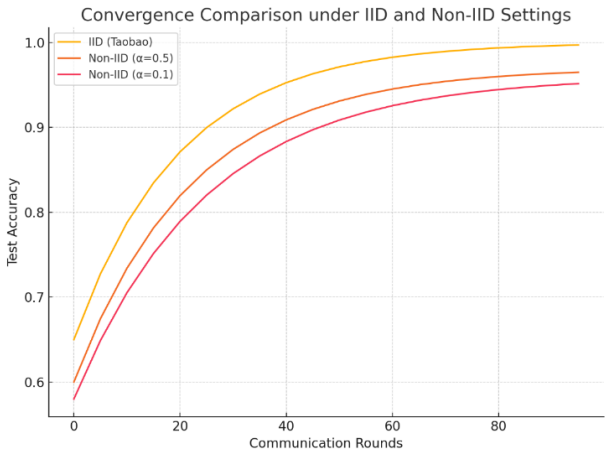

Figure 12. Convergence comparison under IID and non-IID settings.

Besides, given the fast-paced advancements in both FL and LLMs, the potential for impactful applications in the advertising industry is high. This research serves as a stepping stone for future developments that could reshape the landscape of privacy-preserving advertising analytics, with potential applications extending beyond advertising to other industries requiring large-scale, privacy-sensitive data processing.

References

[1] Bengio Y., Ducharme R., Vincent P., and Jauvin C., "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, vol. 3, pp. 1137-1155, 2003. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

[2] Brown T., Mann B., Ryder N., Subbiah M., and *et al.*, "Language Models are Few-Shot Learners," *in Proceedings of the 34th Conference on Neural Information Processing Systems*, Vancouver, pp. 1877-1901, 2020. https://api.semanticscholar.org/CorpusID:218971783

[3] DeepSeek-AI, DeepSeek-V3, Technical Report,

2025. https://arxiv.org/pdf/2412.19437

[4] Gai K., Zhu X., Li H., Liu K., and Wang Z., "Learning Piece-Wise Linear Models from Large Scale Data for Ad Click Prediction," *arXiv Preprint*, vol. arXiv:1704.05194, pp. 1-12, 2017. https://doi.org/10.48550/arXiv.1704.05194

[5] Gentry C., "Fully Homomorphic Encryption Using Ideal Lattices," *in Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, Bethesda, pp. 169-178, 2009. https://doi.org/10.1145/1536414.1536440

[6] Giray L., "Prompt Engineering with ChatGPT: A Guide for Academic Writers," *Annals of Biomedical Engineering*, vol. 51, no. 12, pp. 2629-2633, 2023. https://doi.org/10.1007/s10439-023-03272-4

[7] Hani A., Tagougui N., and Kherallah M., "Toward Human-Level Understanding: A Systematic Review of Vision-Language Models for Image Captioning," *The International Arab Journal of Information Technology*, vol. 23, no. 1, pp. 81-97, 2026. DOI:10.34028/iajit/23/1/8

[8] Kairouz P. and McMahan H., "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1-2, pp. 1-210, 2021. https://doi.org/10.1561/2200000083

[9] Li T., Sahu A., Talwalkar A., and Smith V., "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50-60, 2020. doi:10.1109/MSP.2020.2975749

[10] Li T., Sahu A., Zaheer M., Sanjabi M., and et al., "Federated Optimization in Heterogeneous Networks," *in Proceedings of the Conference on Machine Learning and Systems*, Texas, pp. 429-450, 2020. https://api.semanticscholar.org/CorpusID:59316566

[11] Li Z., Hou Z., Liu H., Li T., and et al., "Federated Learning in Large Model Era: Vision-Language Model for Smart City Safety Operation Management," *in Proceedings of the Companion Proceedings of the ACM Web Conference*, pp. 1578-1585, 1578-1585, 2024. https://doi.org/10.1145/3589335.3651939

[12] McMahan B., Moore E., Ramage D., Hampson S., and Arcas B., "Communication-Efficient Learning of Deep Networks from Decentralized Data," *in Proceedings of the 20th International Conference Artificial Intelligence and Statistics*, Florida, pp. 1273-1282, 2017. https://scispace.com/pdf/communication-efficient-learning-of-deep-networks-from-2s16evj791.pdf

[13] Radford A., Narasimhan K., Salimans T., and Sutskever B., Improving Language Understanding by Generative Pre-Training, OpenAI, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, Last Visited, 2025.

[14] Radford A., Wu J., Child R., Luan D., and et al., Language Models are Unsupervised Multitask Learners, OpenAI Blog, https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf. Last Visited, 2025.

[15] Roth H., Zephyr M., and Harouni A., Federated Learning with Homomorphic Encryption, NVIDIA Developer Blog, https://developer.nvidia.com/blog/federated-learning-with-homomorphic-encryption, Last Visited, 2025.

[16] Tianchi, Ad Display/Click Data on Taobao.com, Alibaba Cloud Tianchi, https://tianchi.aliyun.com/dataset/dataDetail?dataId=56, Last Visited, 2025.

[17] Vaswani A., Shazeer, N., Parmar N., Uszkoreit J., and et al., "Attention is all you Need," *in Proceedings of the 31st International Conference on Neural Information Processing Systems*, California, pp. 5998-6008, 2017. https://doi.org/10.48550/arXiv.1706.03762

[18] White J., Fu Q., Hays S., Sandborn M., and et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," *arXiv Preprint*, vol. arXiv:2302.11382v1, pp. 1-19., 2023. https://doi.org/10.48550/arXiv.2302.11382

[19] Xie Q., Jiang S., Jiang L., Huang Y., and et al., "Efficiency Optimization Techniques in Privacy-Preserving Federated Learning with Homomorphic Encryption: A Brief Survey," *IEEE Internet Things Journal*, vol. 11, no. 14, pp. 24569-24580, 2024. DOI:10.1109/JIOT.2024.3382875

[20] Zhang J., Yang H., Li A., Guo X., and et al., "MLLM-FL: Multimodal Large Language Model Assisted Federated Learning on Heterogeneous and Long-Tailed Data," *arXiv Preprint*, vol. arXiv:2409.06067v2, pp. 1-11, 2024. https://doi.org/10.48550/arXiv.2409.06067

[21] Zhou G., Song C., Zhu X., Ying Fan., and et al., "Deep Interest Network for Click-Through Rate Prediction," *arXiv Preprint*, vol. arXiv:1706.06978v4, pp. 1-9, 2017. https://doi.org/10.48550/arXiv.1706.06978

Jialu Li is an undergraduate student at the School of Mathematics and Statistics, Central South University. She has participated in a number of scientific research competitions and social practice programs. Her research interests focus on Large Language Models, Data Analysis, AI+Education, Financial Technology, and Quantitative modeling.