Marketing Analytics Data Scientist Case Study

Introduction

An e-commerce site sells through various channels. The aim of this platform is to analyze user events from different channels. On this e-commerce site, all events are available from the moment the user first enters the site to the moment they log out.

Using the data sources collected from BigQuery, it is aimed at determining how we can move forward to increase the engagement of the platform.

The aim is to explore how we can keep users on this platform and, at the same time, increase the sales of the e-commerce site. Also, various data analysis, machine learning, and visualization libraries will be used to analyze the data. Diverse ideas and models will then be generated to increase user engagement.

Methodology

- Data Collection & Processing
    - Data Extraction via BigQuery
    - Editing Time Columns on Data
    - Finding Empty Values on Data
- Data Analysis And Visualization
    - Making Sense of Columns on Data
    - Visualizing Meaningful Data
    - Commenting on Data Visualizations
- Insight and Recommendation Models
    - Thinking of New Models and Ideas
    - Interpretation of Models
- Integration with Flow

Data Collection And Processing

Data Collection & Processing

First of all, in the Data Collection step, the data related to the website is extracted from BigQuery with an assigned BigQuery query. At this stage, several queries are written on BigQuery to find the required data in the healthiest way. At the end, all columns that can be used with the query that gives the best result are retrieved.

After this stage, we import the data and libraries we will use. As a result of the phase, it is decided how to group the data in the data visualization phase by looking at the unique values of the data of our columns and finding the missing values in our data.

When we examine the data, we see that a unique value is assigned for each user. Depending on the unique value of the users, the event operations performed on a user basis can be tracked. Here, what the user did with which event can be understood with the data in the time column.

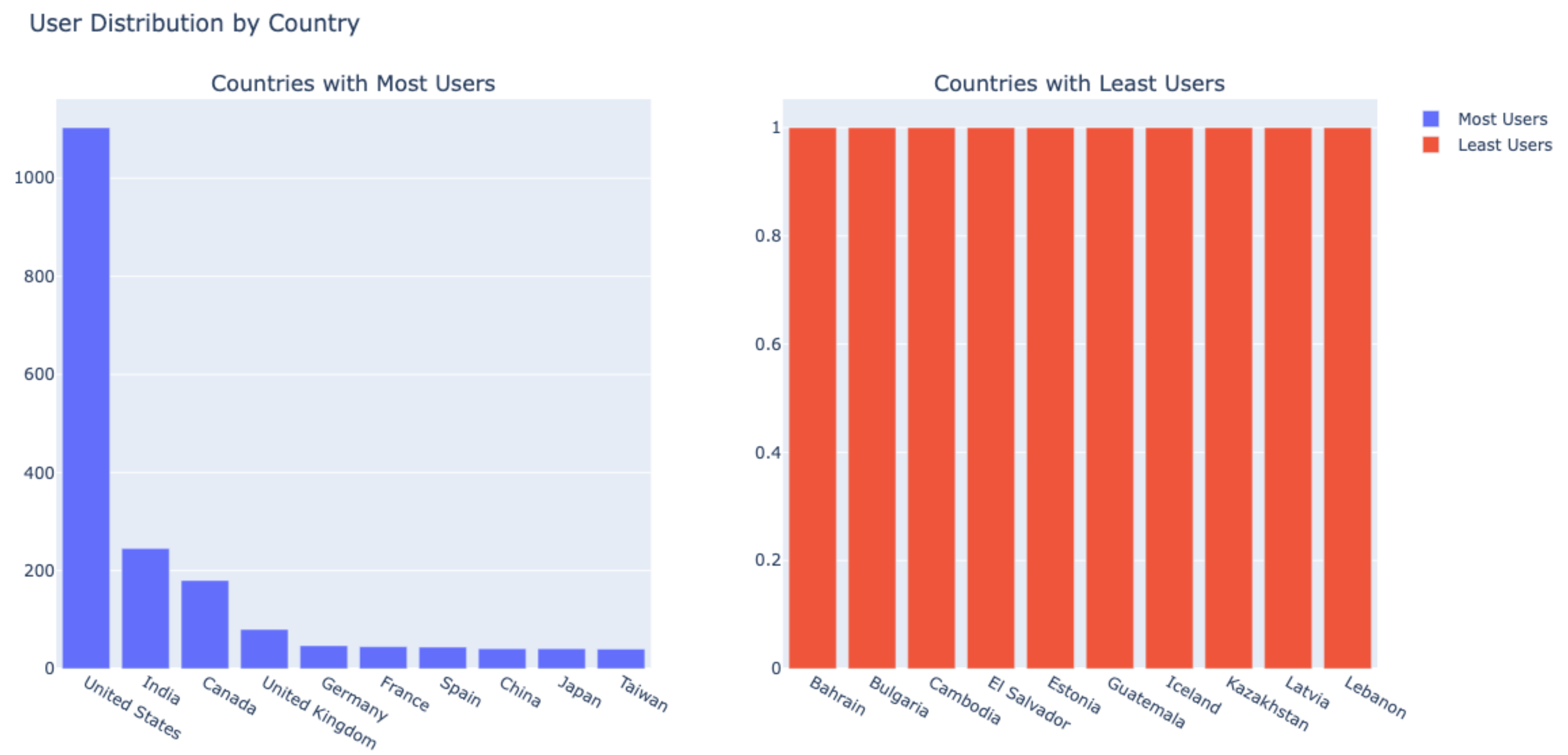
Data Analysis And Visualization

Total User Logins by Country

One of the first things we need to explore in our dataset is the volume of users from different countries. In the charts below, you can see the top 5 countries with the highest and lowest user numbers.

These charts show that the countries with the highest user inflow are the United States, India and Canada. On the other hand, when we look at the countries with the lowest user inflow, we see Estonia, Latvia and Guatemala.

It is important to note that while this information provides valuable insight into countries with significant user engagement, it may not be sufficient for a comprehensive analysis. Further research is needed to gain a deeper understanding of the focal marketplaces and sources of transaction volumes.

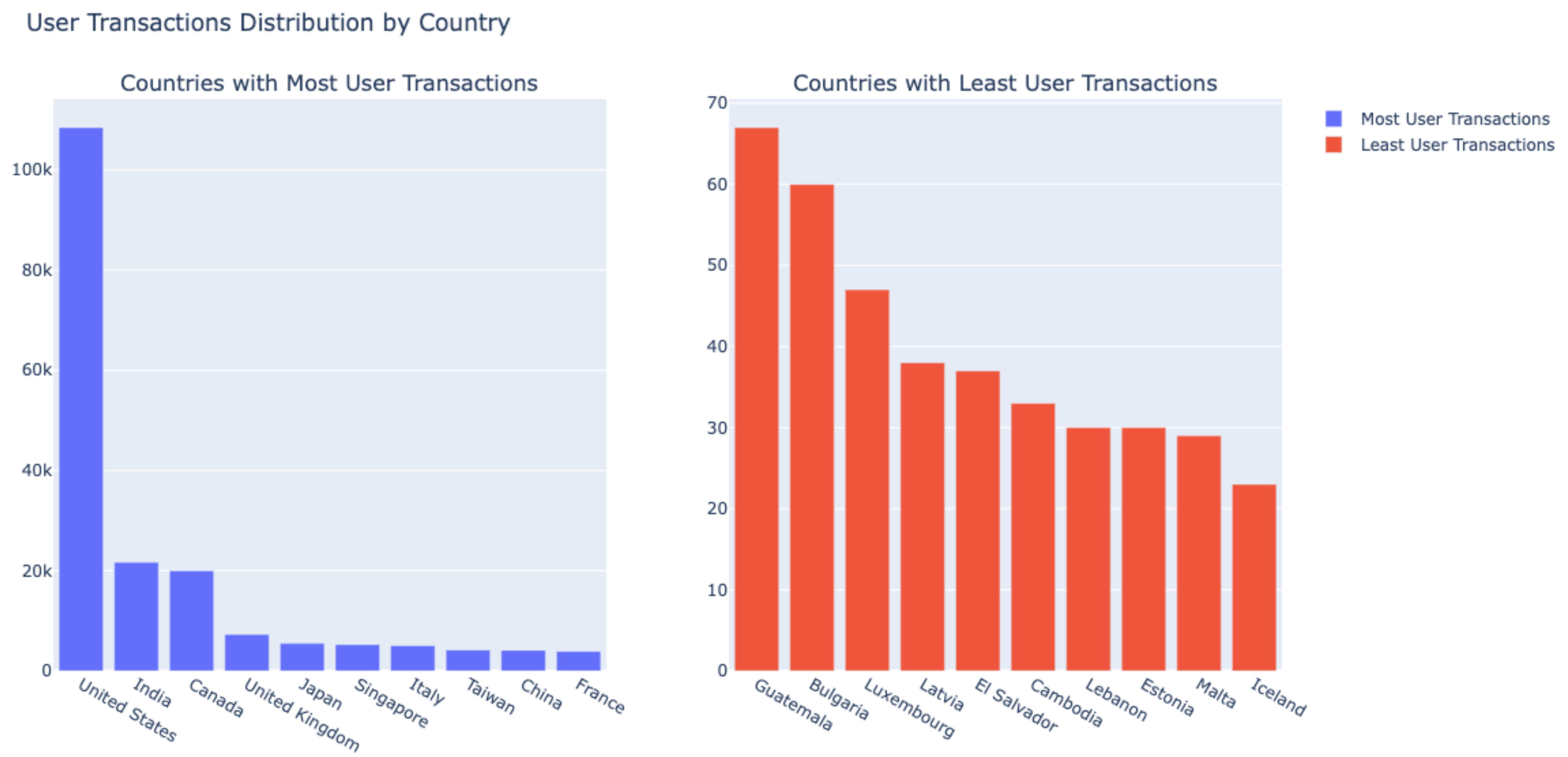


Total User Transaction by Country

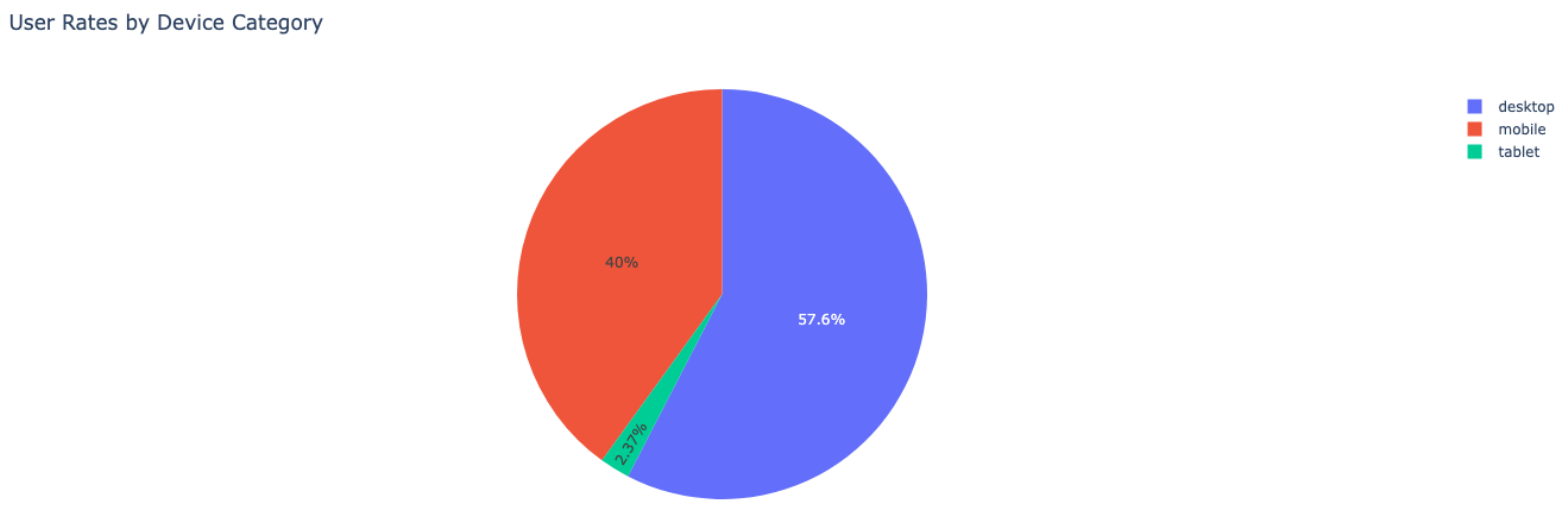
One of the first things we need to explore in our dataset is the volume of users and trades from different countries. In the charts below, you can see the 5 countries with the highest and lowest trading volumes.

These charts show that the countries with the highest user transactions are the United States, India and Canada. On the other hand, when we look at the countries with the lowest user inflow, we see Guatemala, Bulgaria and Luxembourg.

While this information provides valuable insights into countries with significant user engagement, it is important to note that they may not be sufficient for a comprehensive analysis. Further research is needed to gain a deeper understanding of the focal marketplaces and sources of transaction volumes.

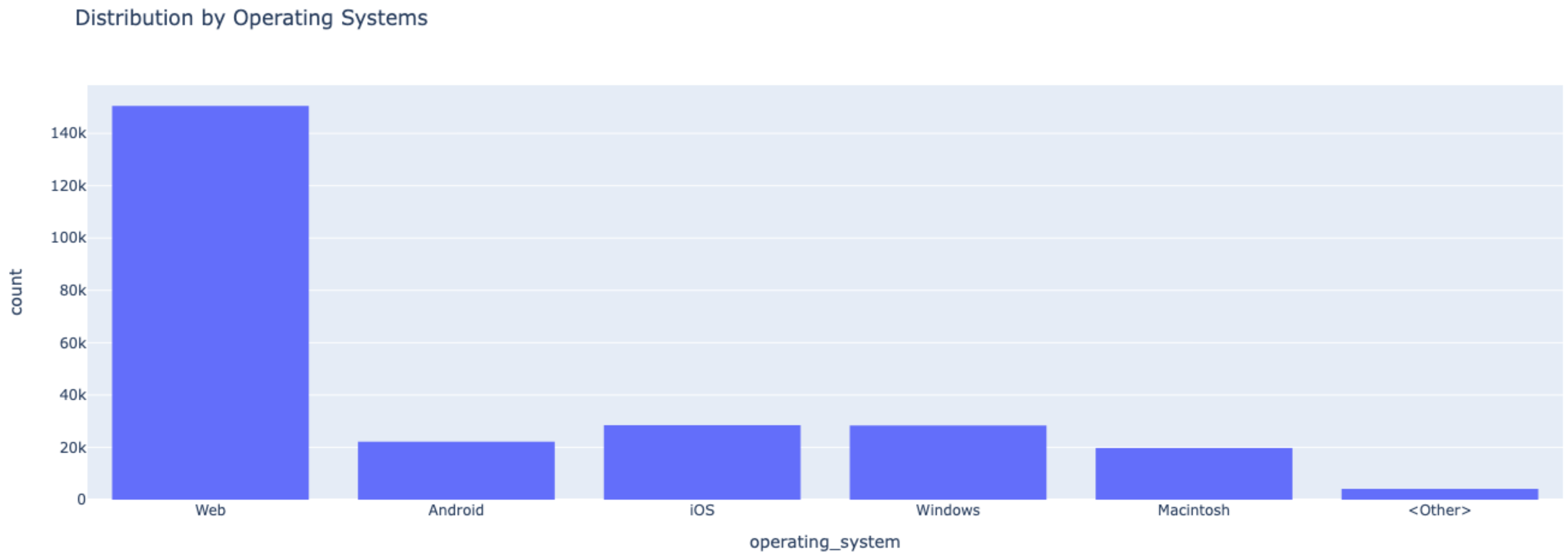


User Rates by Device Category



When we examine the events of the users, it is seen that 58% of the users conducted the events on desktop, 40% on mobile devices and 2% on tablets.

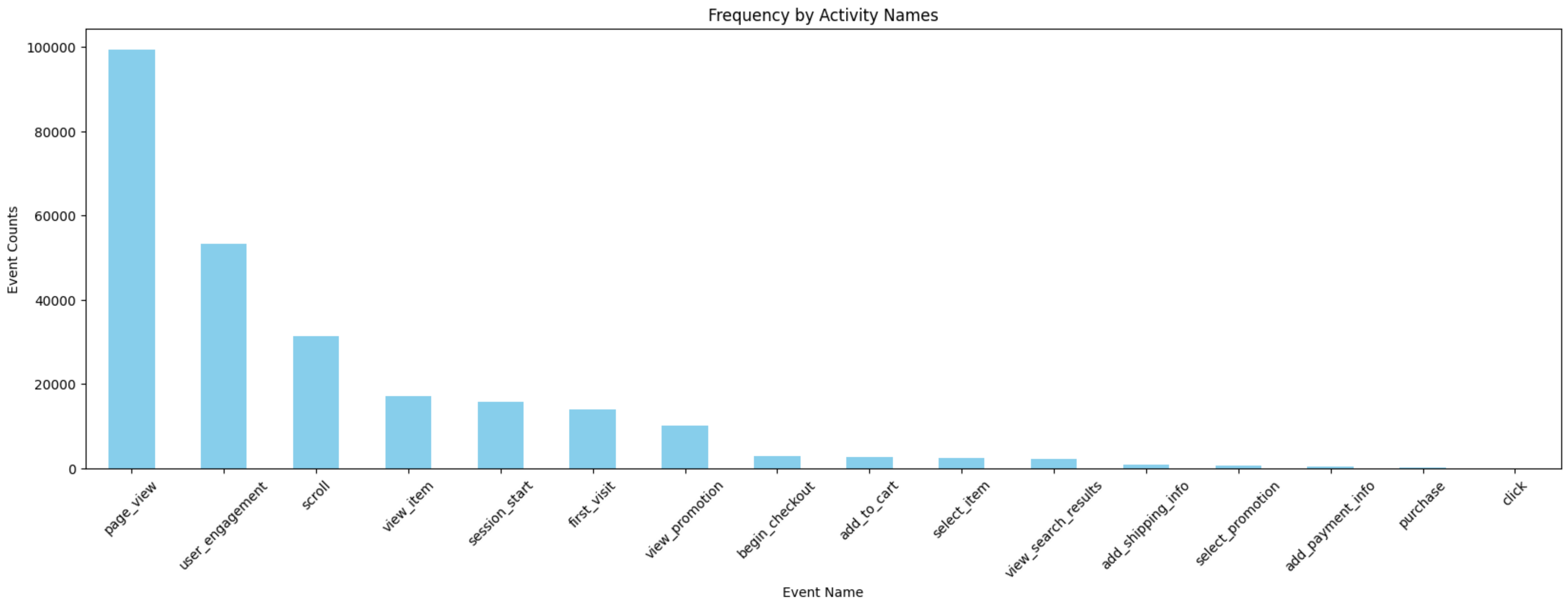
Distribution by Operating Systems



According to the graph, the web category has the highest number of uses, at around 140,000. Among other operating systems, "Android" and "iOS" have moderate usage rates, while "Windows" and "Macintosh" have lower usage rates.

Frequency by Activity Names

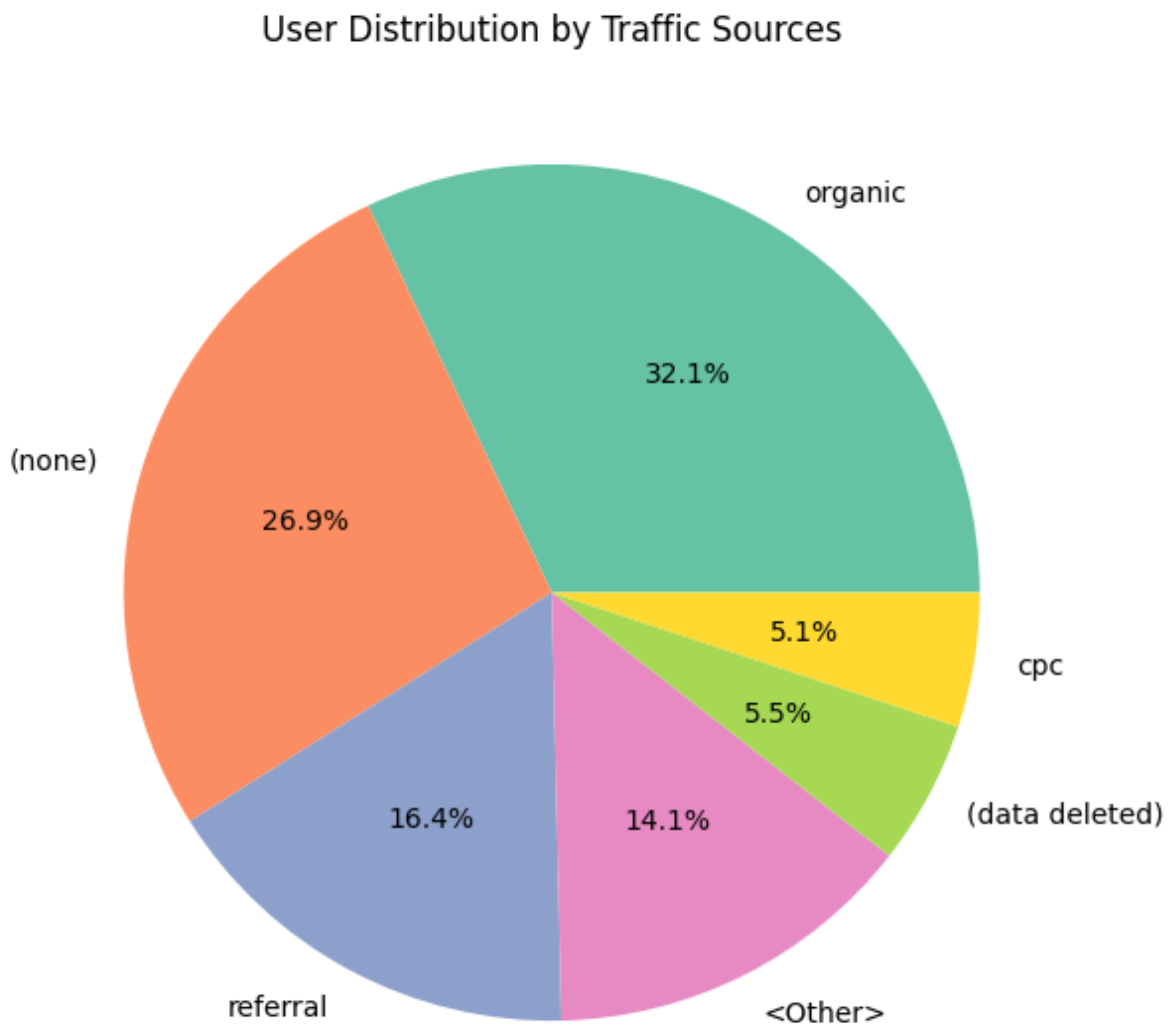
Looking at the bar chart, the 'page_view' event appears to have the highest frequency compared to all other events. Events include 'user_engagement', 'scroll', 'view_item', 'session_start', 'first_visit', and other e-commerce related events ('add_to_cart', 'purchase', 'select_item', etc.). The 'click' event on the right-hand side of the graph has the lowest number, indicating that users conduct this action the least. The visual serves as an example of a data visualization that can be used to analyze users' event within a website or app.



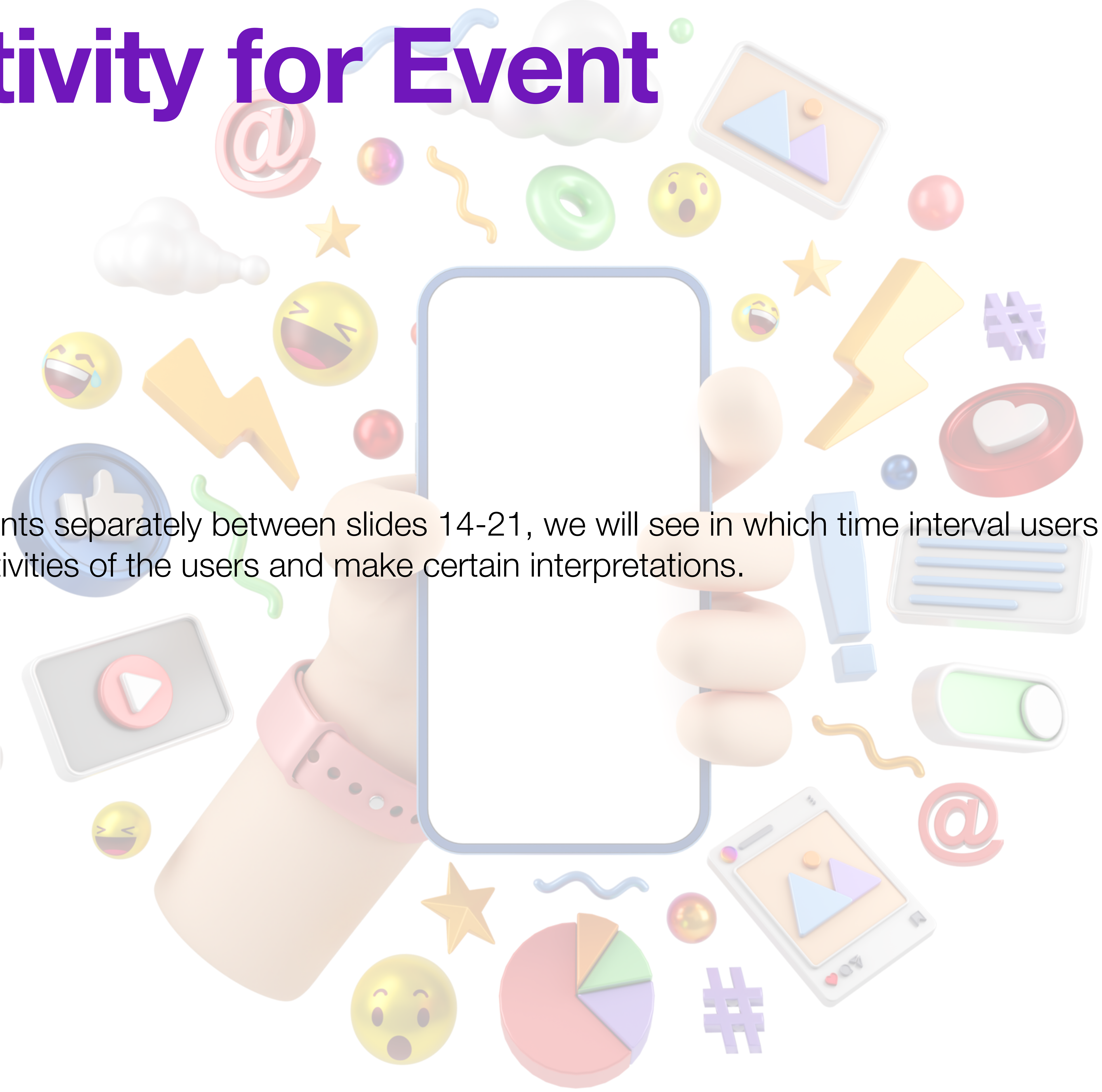
User Distribution by Traffic Sources

Looking at the graph of user distribution by traffic source, the largest share is "organic," i.e., traffic from organic searches, which accounts for 32.1% of total traffic. Visits with no traffic source (none) come in second, representing 26.9%. "Referral" traffic is in third place with 16.4%.

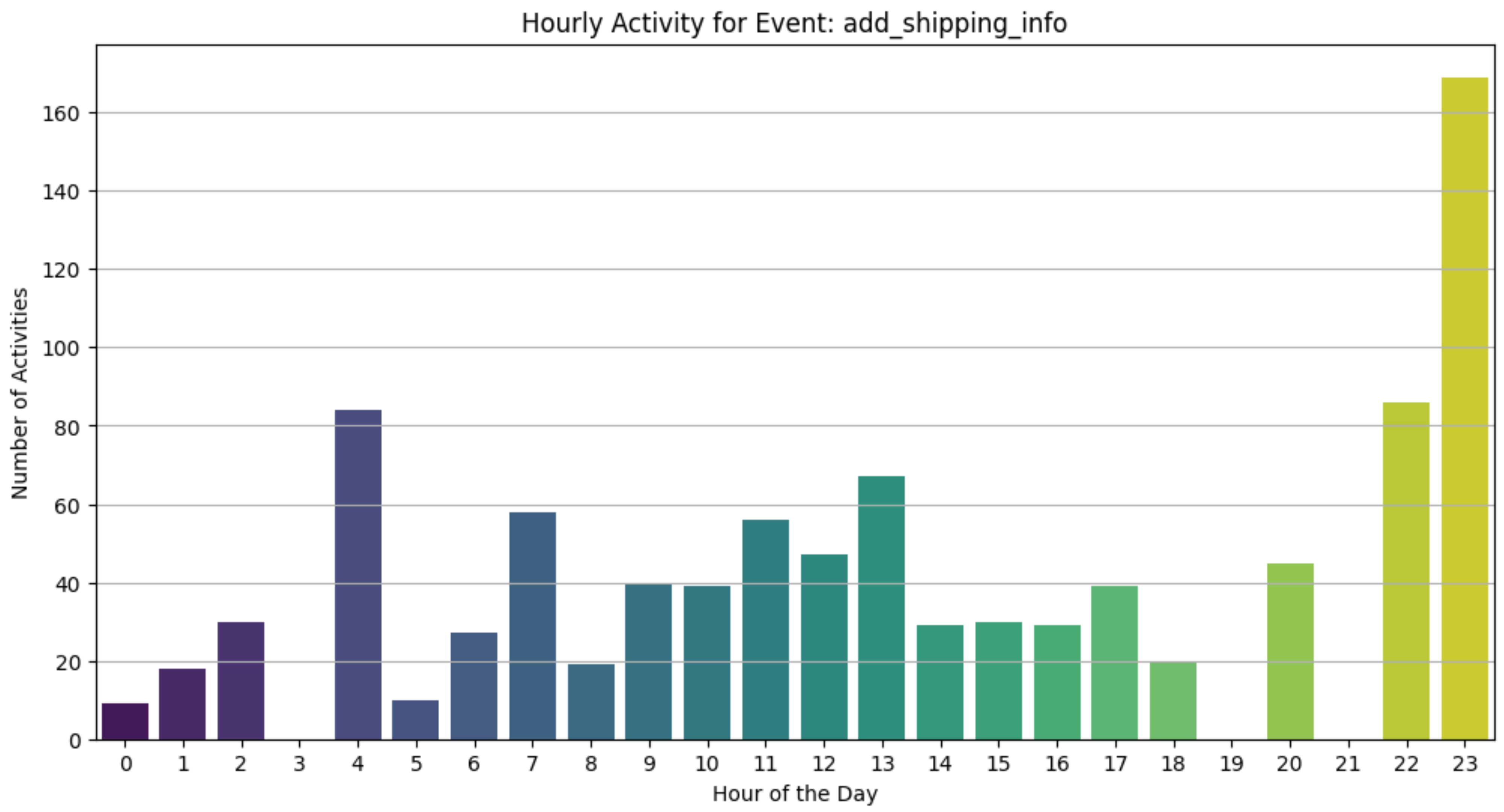
This chart provides useful information for a website owner or marketer, as understanding traffic sources is critical to developing marketing and content strategies.

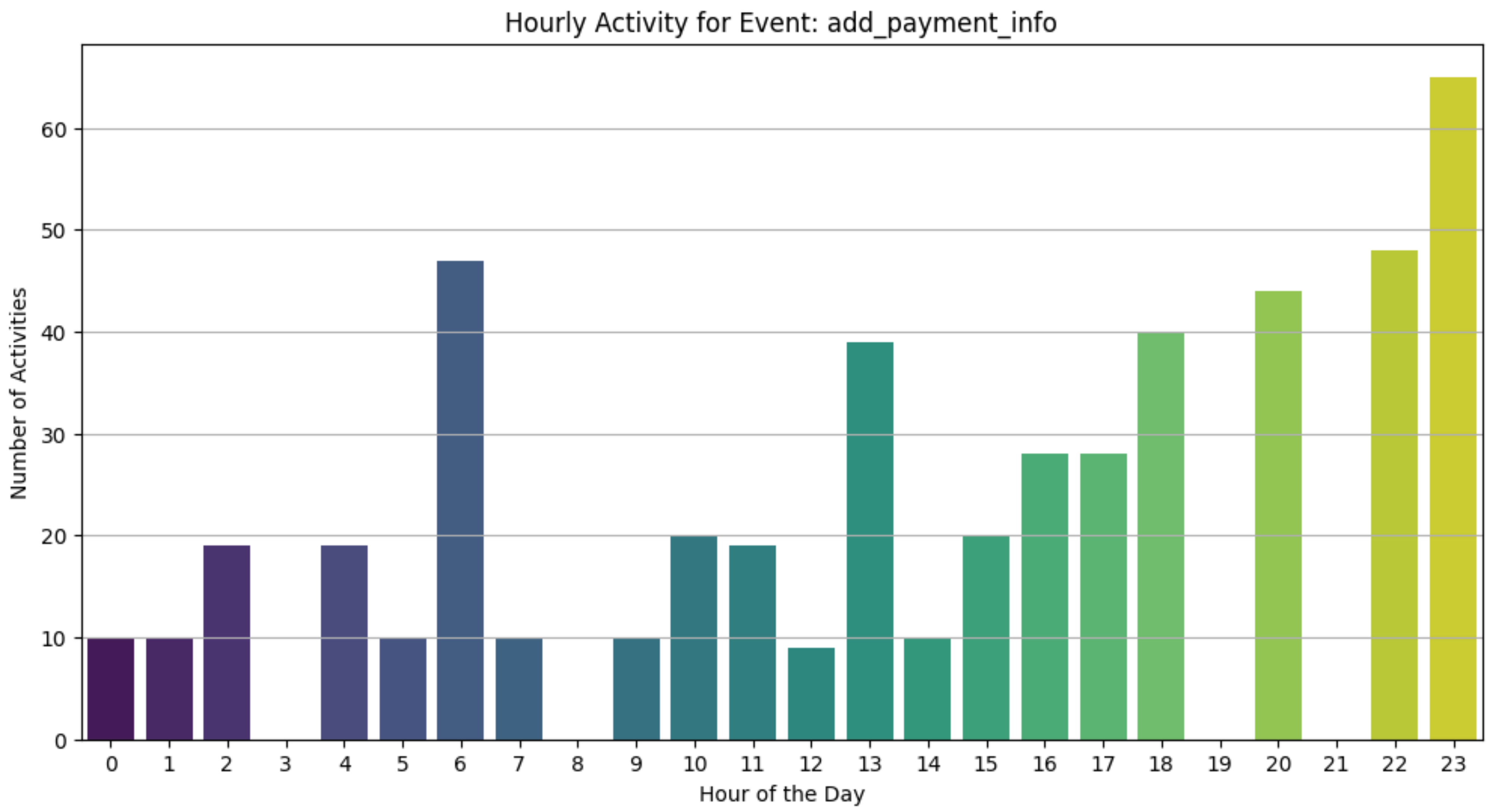


Hourly Activity for Event

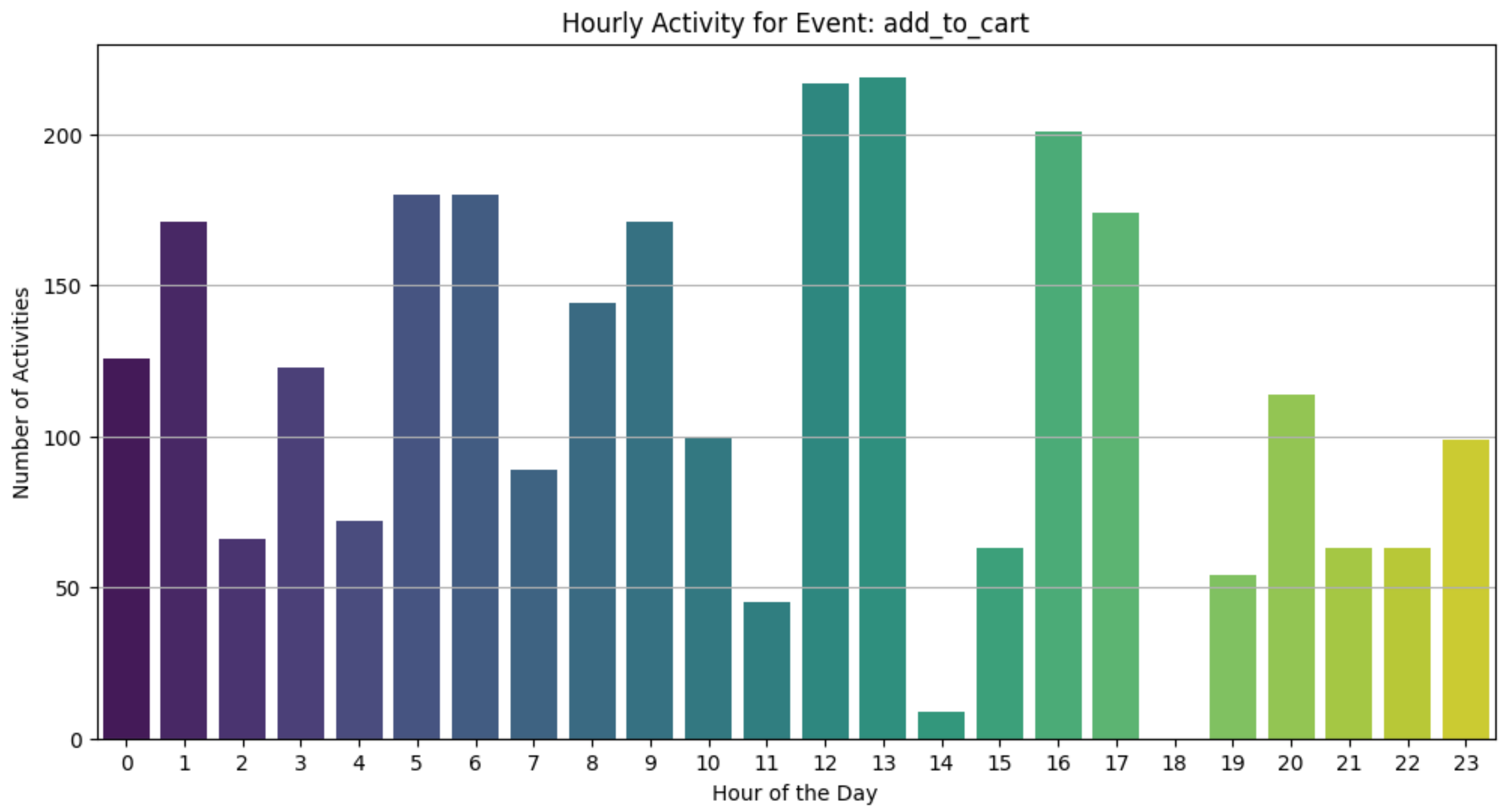


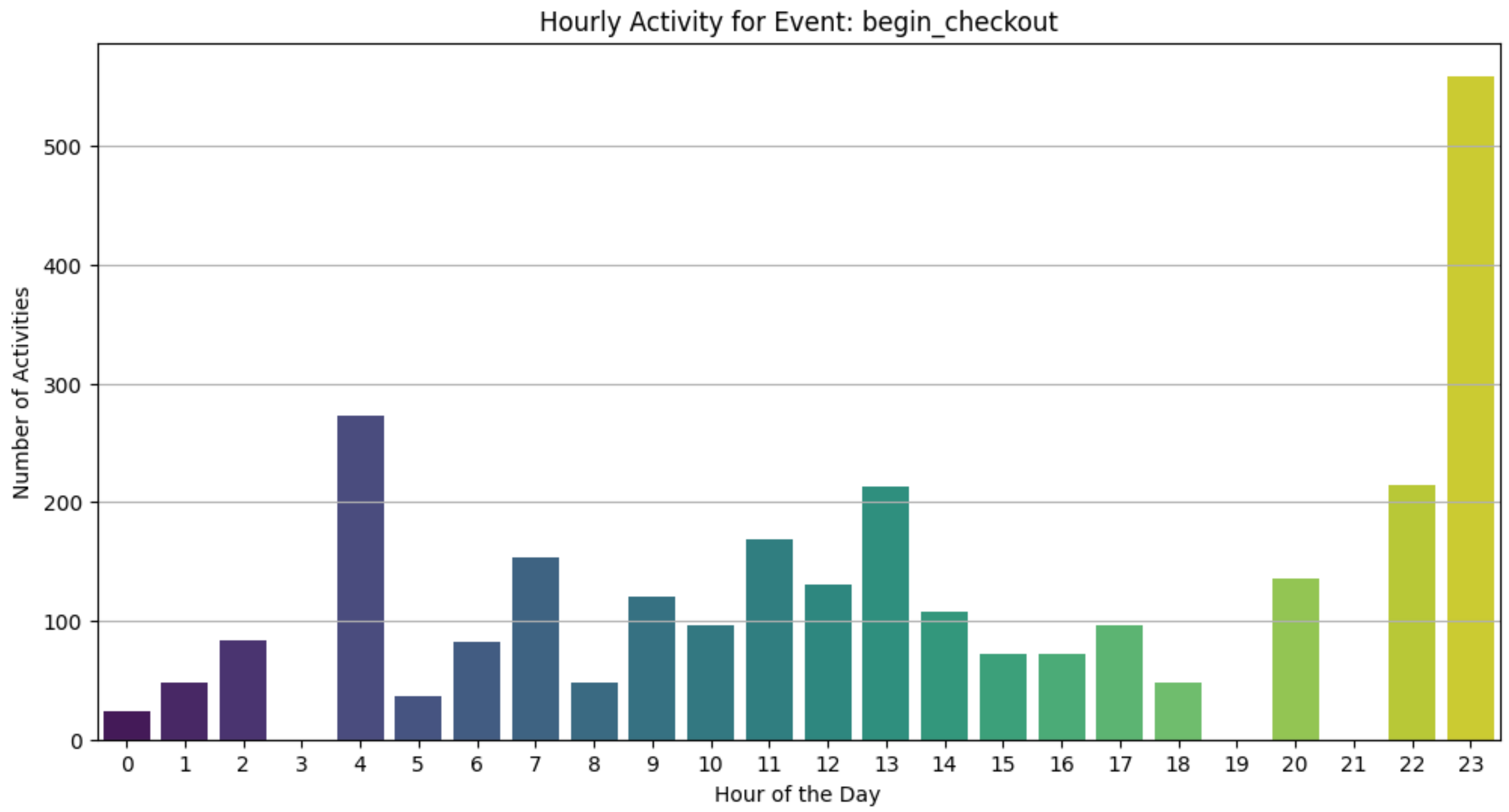
By visualizing each of the user events separately between slides 14-21, we will see in which time interval users use the website more. Afterwards, we will analyze the activities of the users and make certain interpretations.



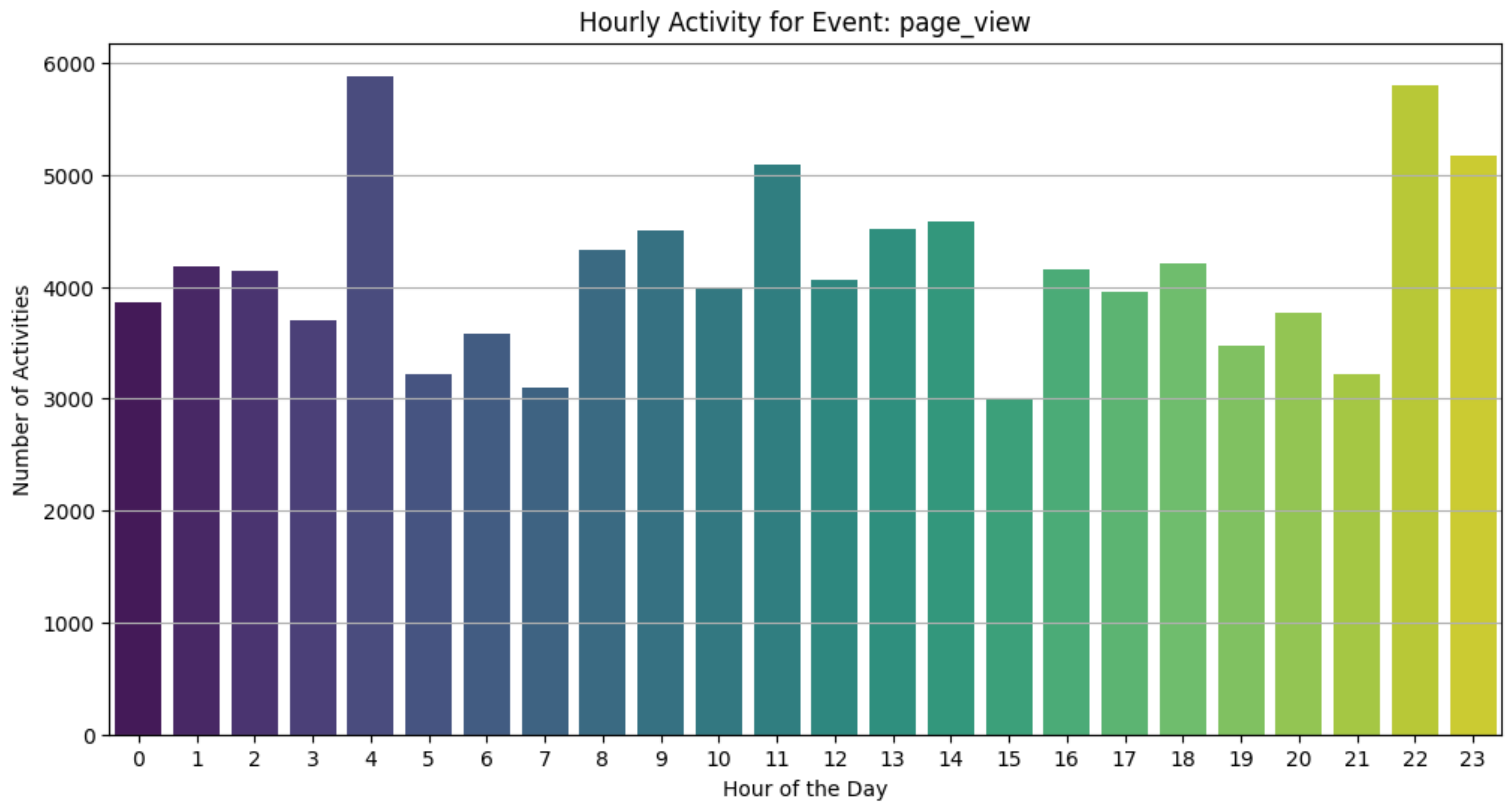


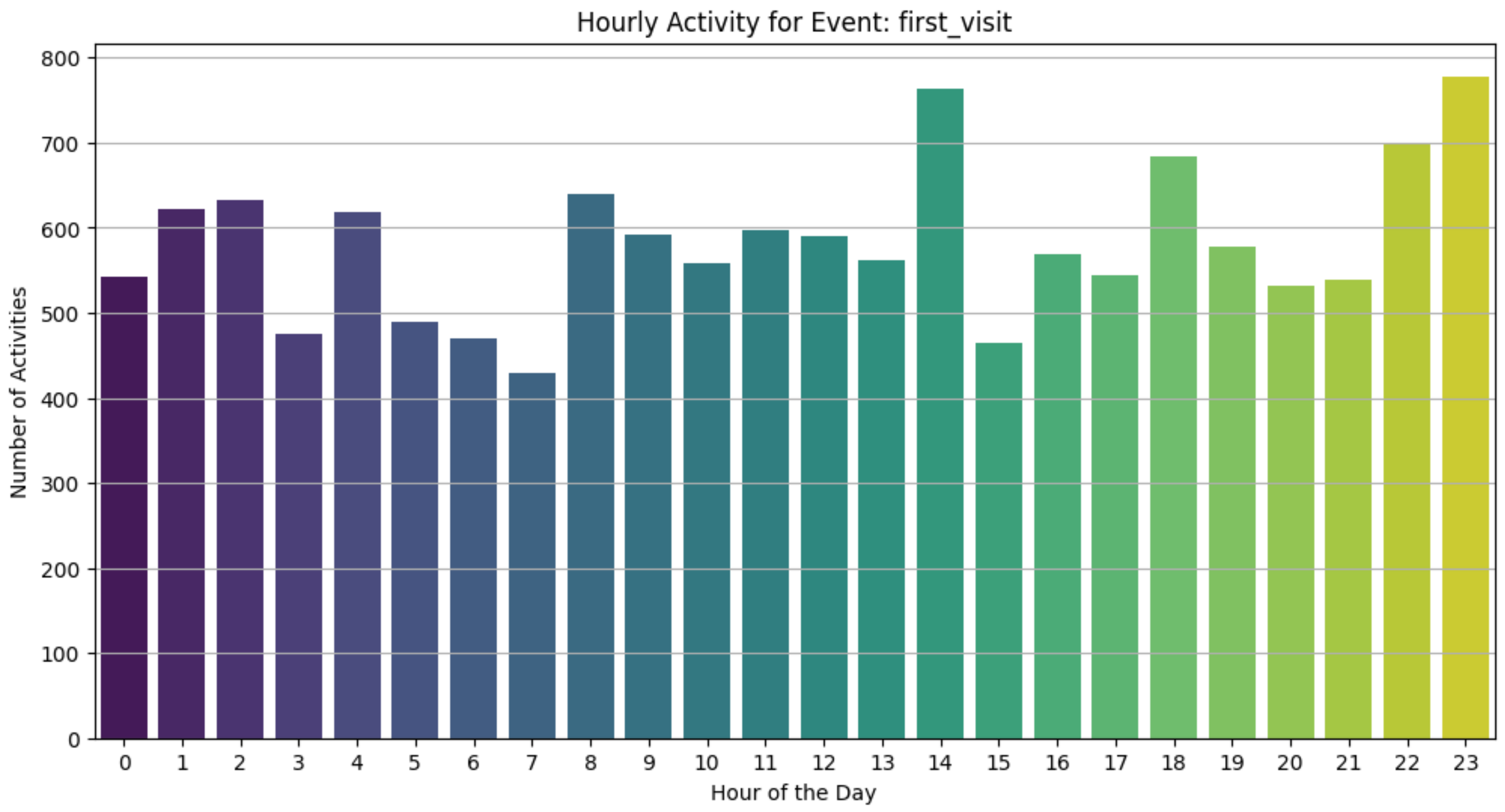
add_shipping_info : Looking at the user events in the graph, it is seen that the Add Shipping Info event conducts most frequently between 18:00 and 00:00.
add_payment_info : Looking at the user events in the chart, it is seen that the Add Payment Info event conducts mostly between 18:00 and 00:00.



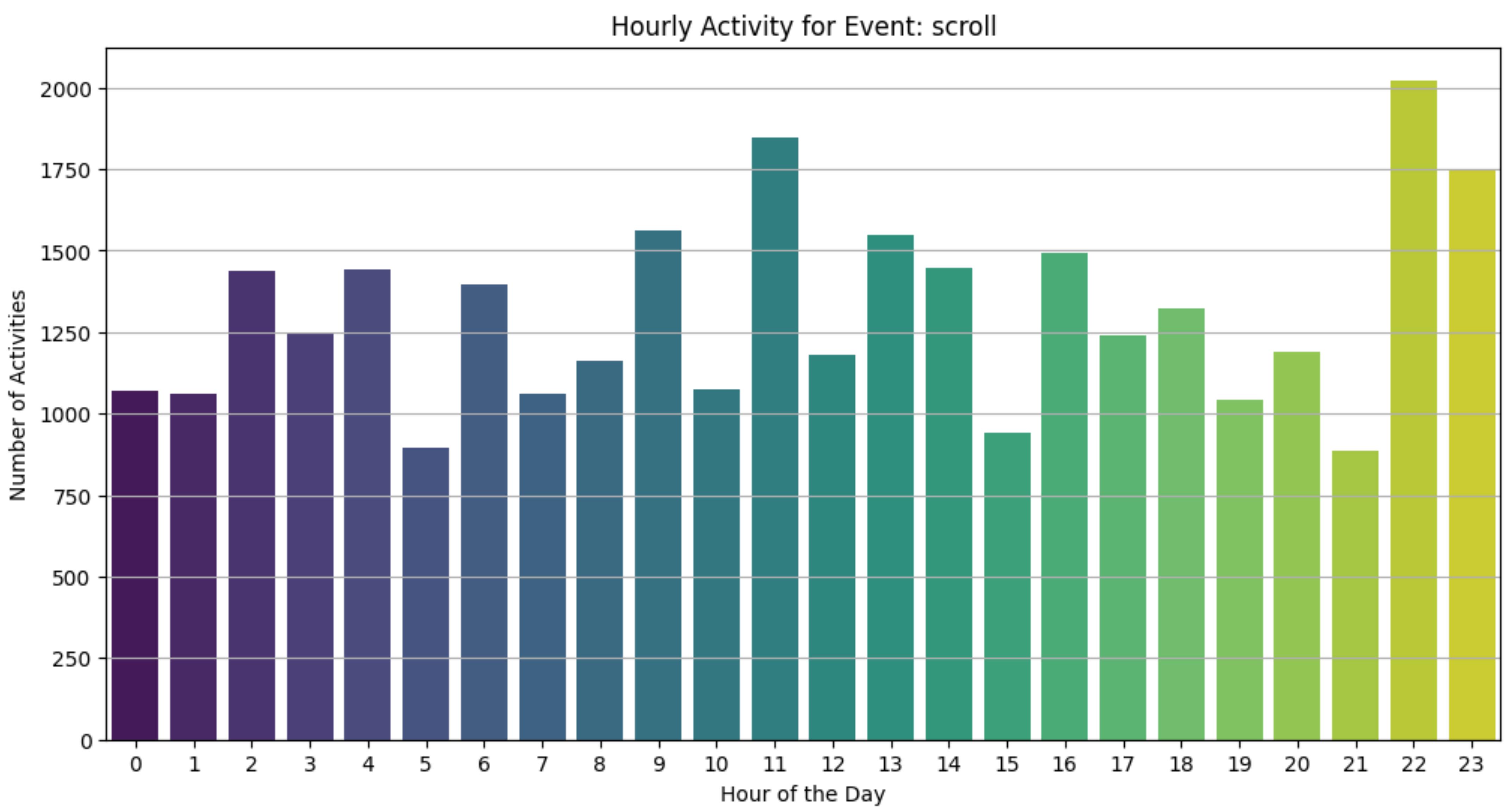


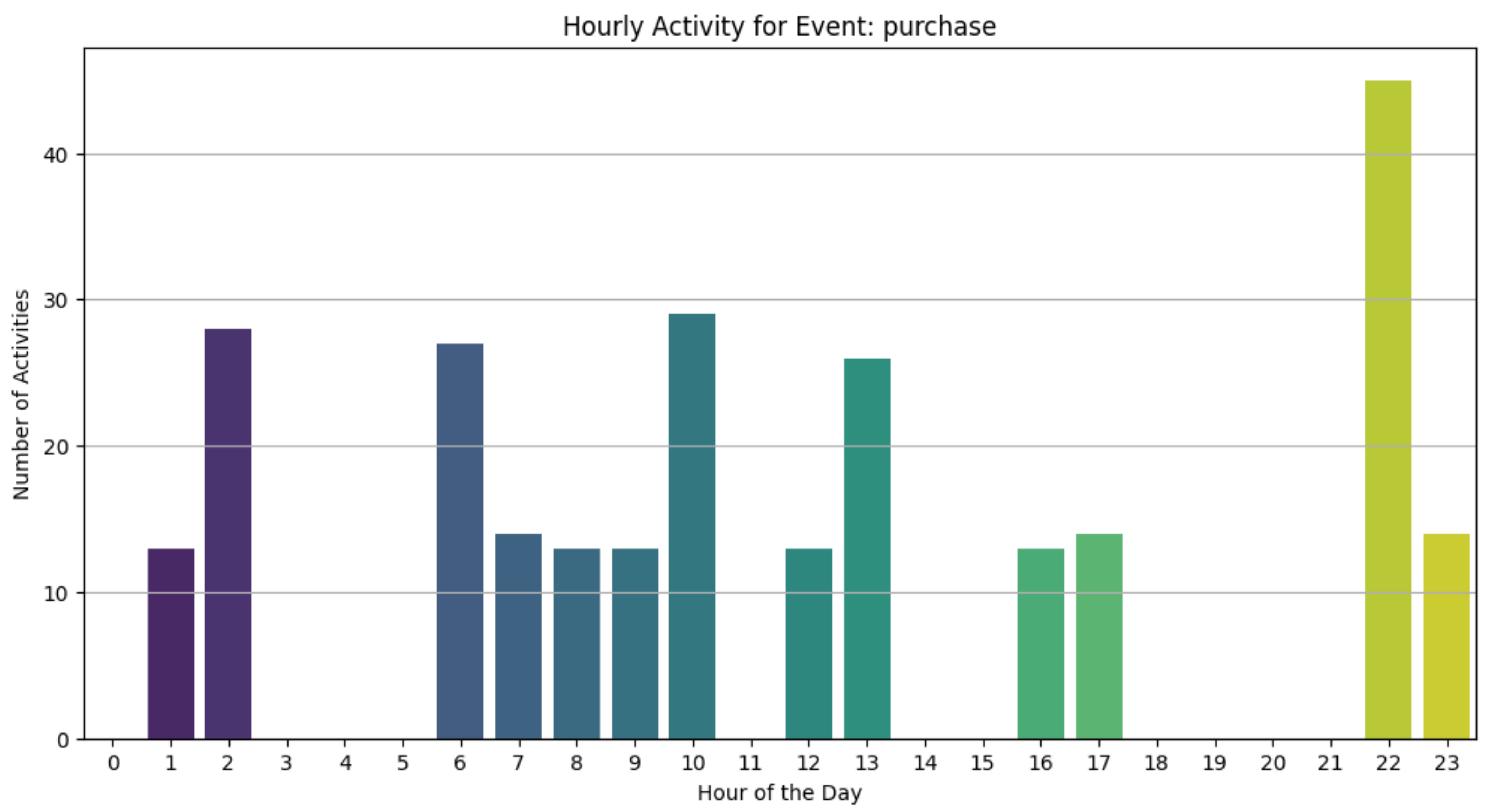
add_to_cart : Looking at the user events in the chart, it is seen that the Add to Cart event conducts mostly between 12:00 and 18:00.
begin_checkout : Looking at the user events in the chart, it is seen that the Begin Checkout event conducts mostly between 18:00 and 00:00.



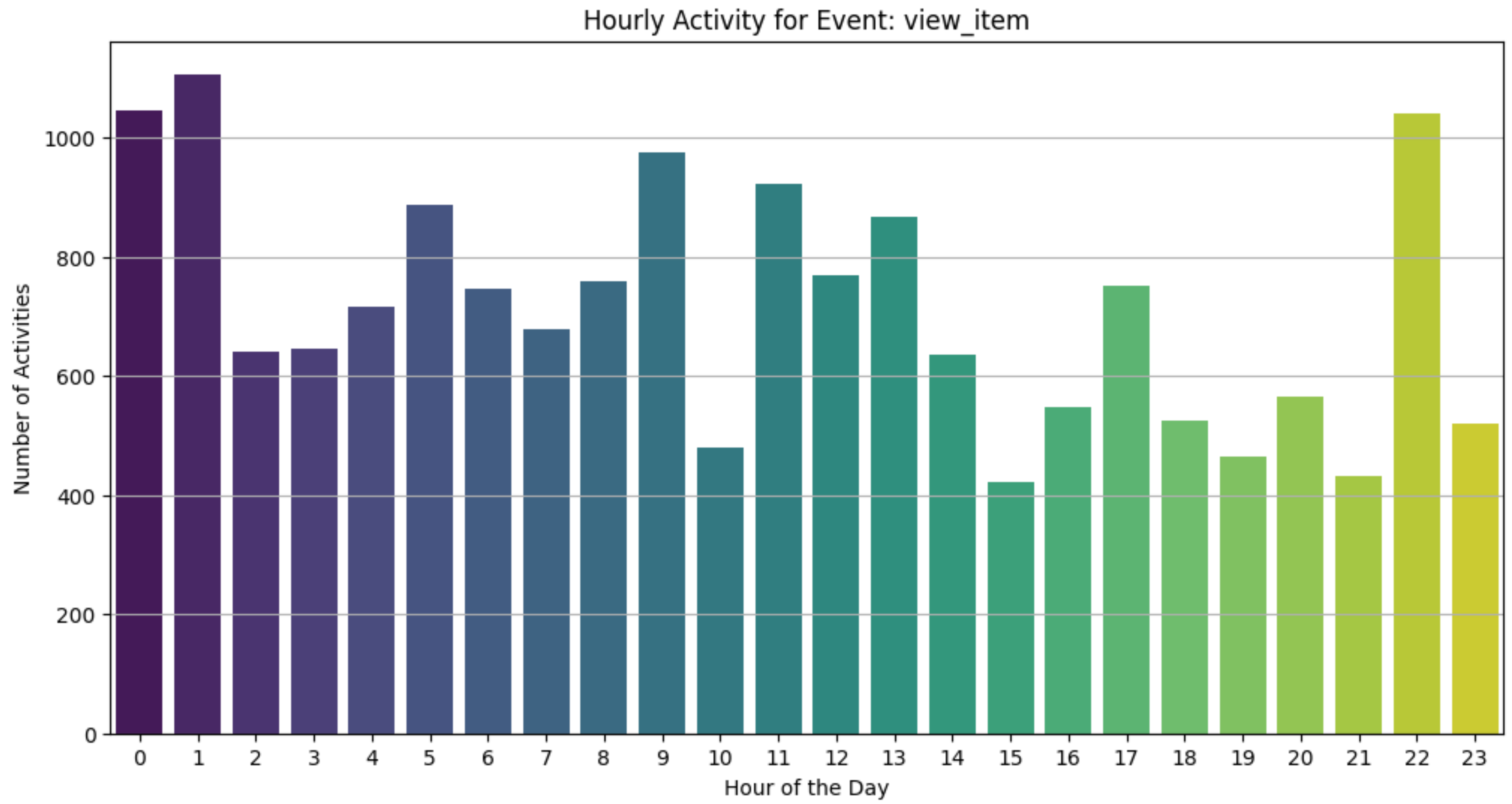


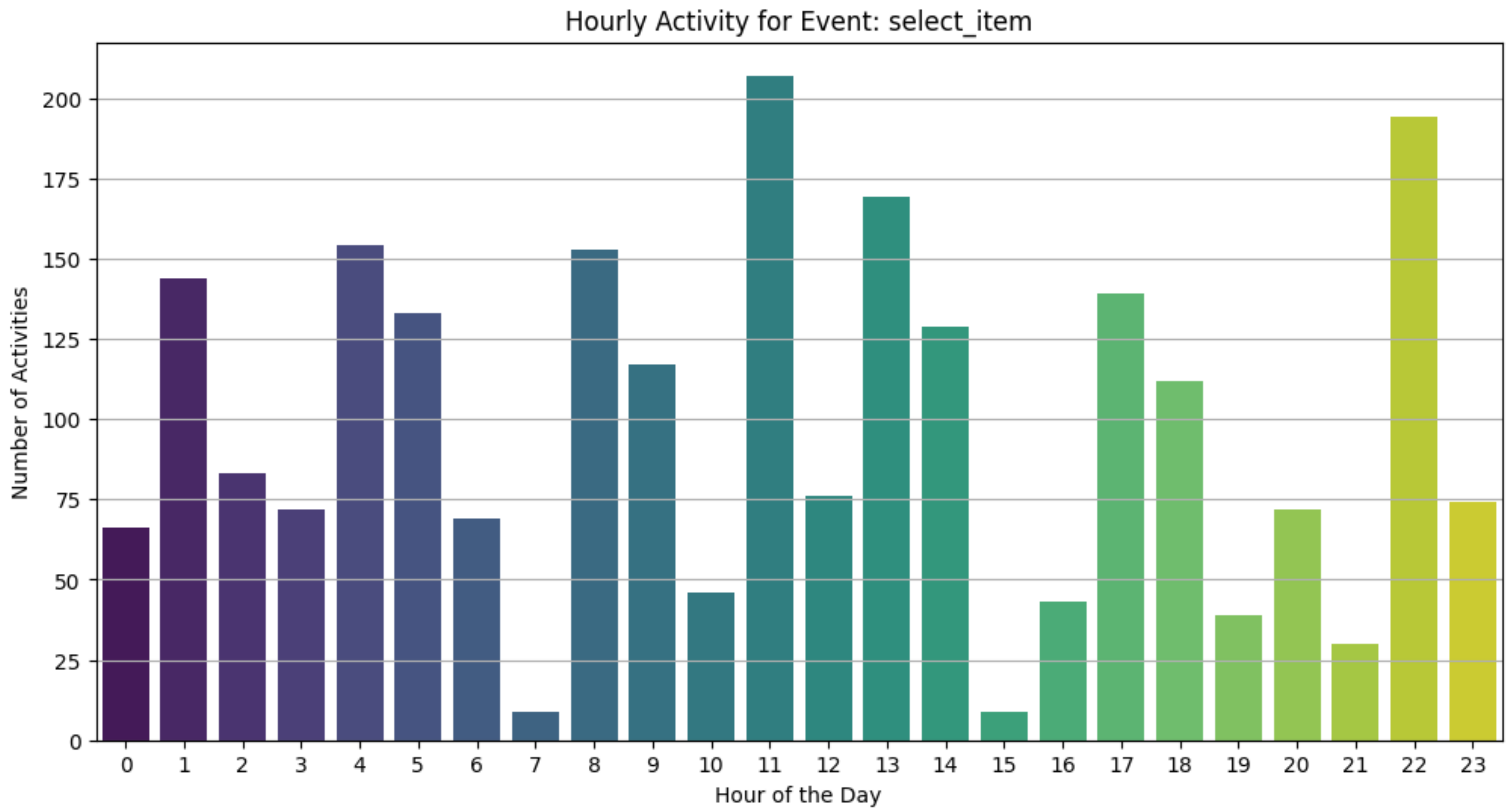
first_visit : Looking at the user events in the chart, it is seen that the First Visit event conducts mostly between 18:00 and 00:00.
page_view : Looking at the user events in the graph, it is seen that the Page View event conducts mostly between 8:00 and 14:00.



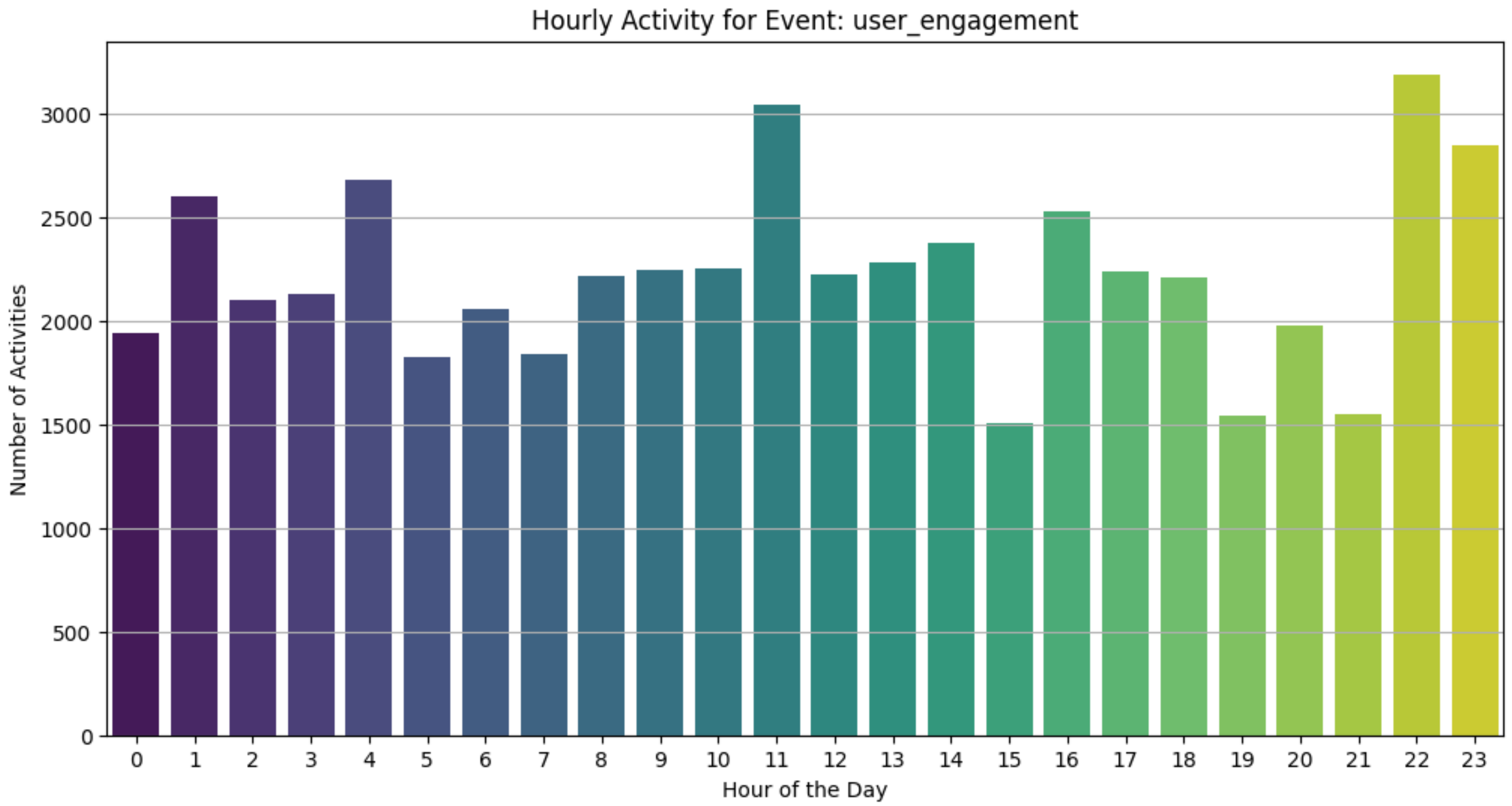


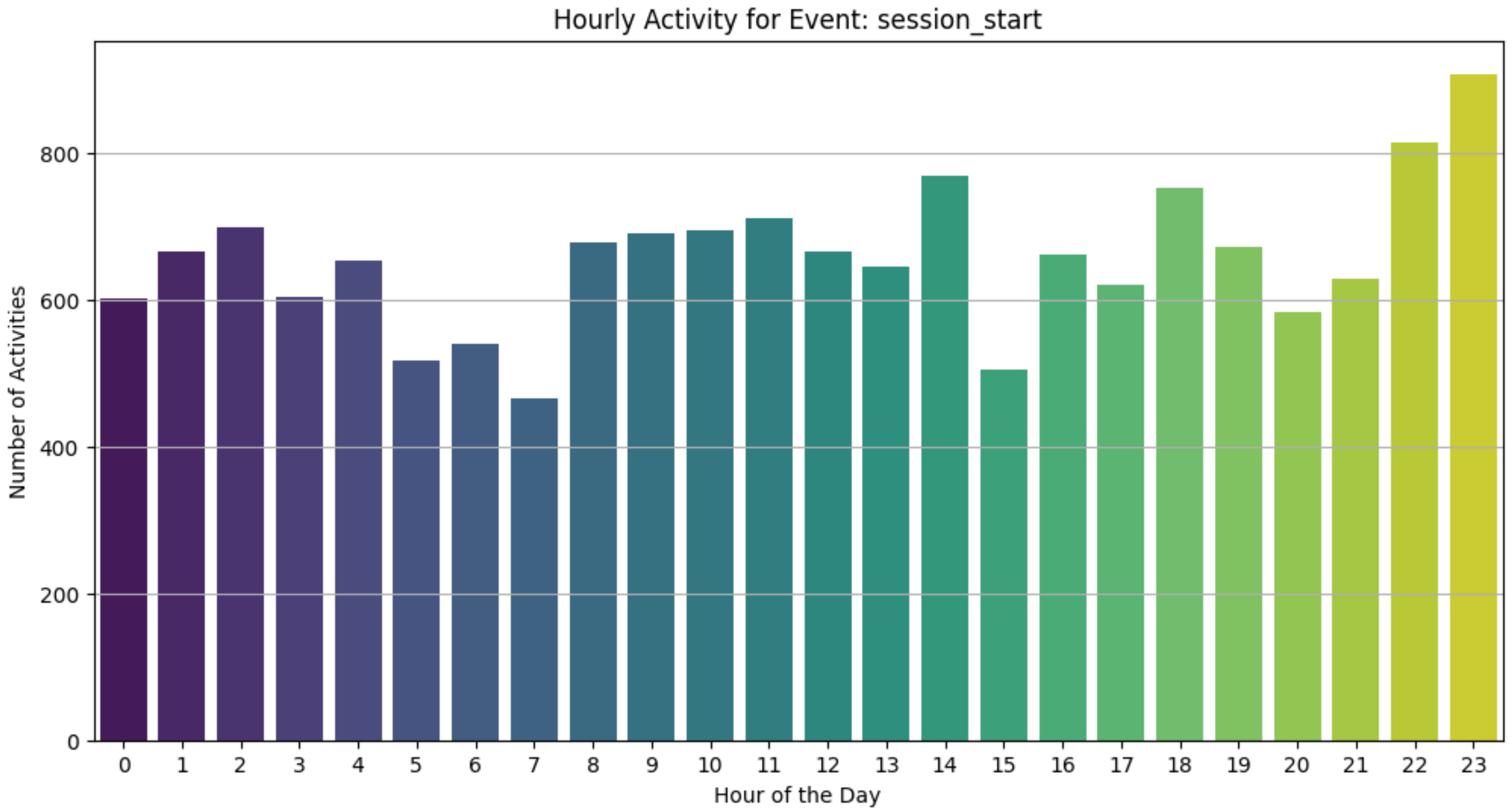
scroll : Looking at the user events in graphic, it is seen that the Scroll event conducts mostly between 18:00 and 00:00.
purchase : Looking at the user events in the chart, it is seen that the Purchase event conducts mostly between 6:00 and 12:00.



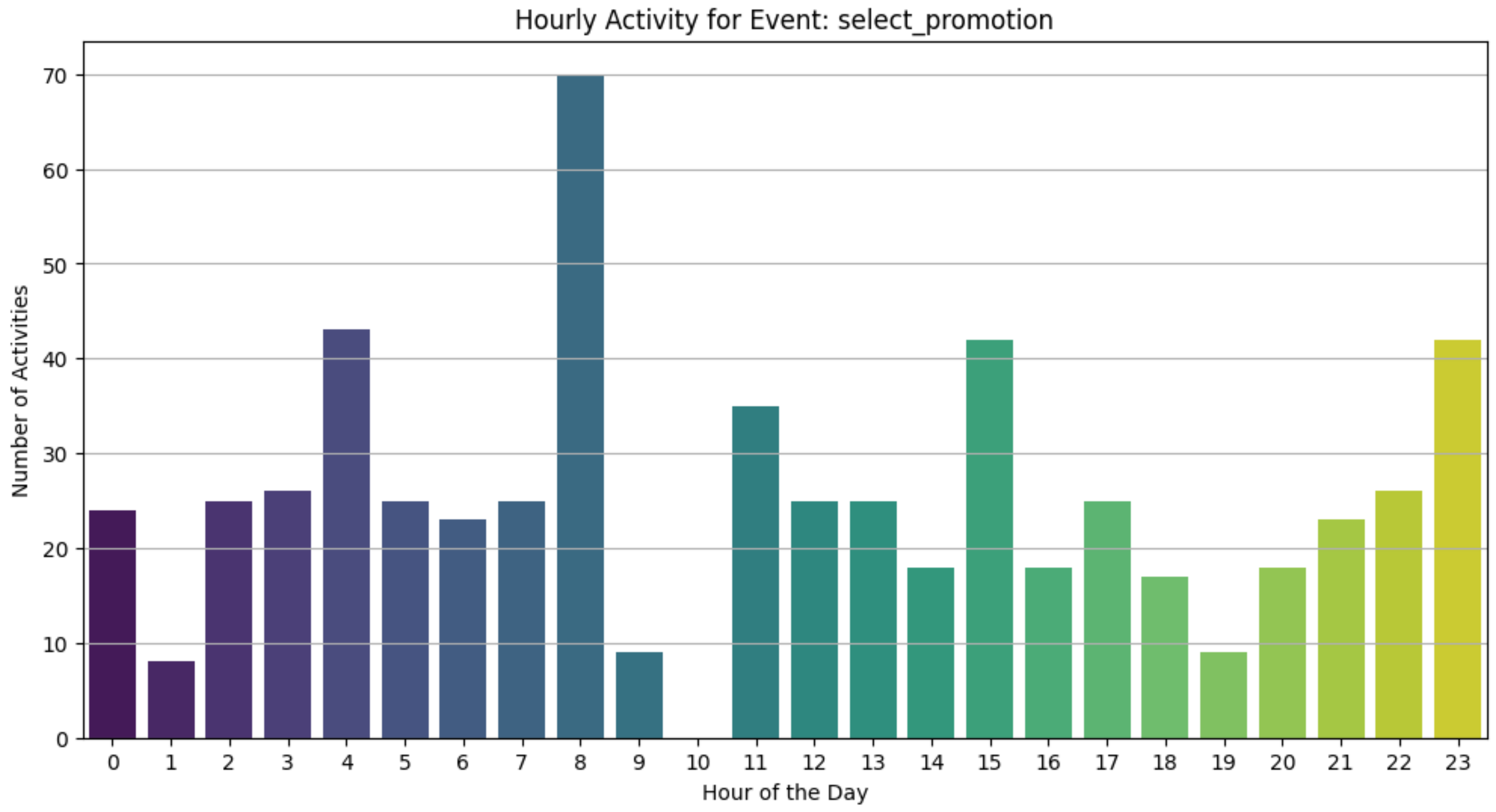


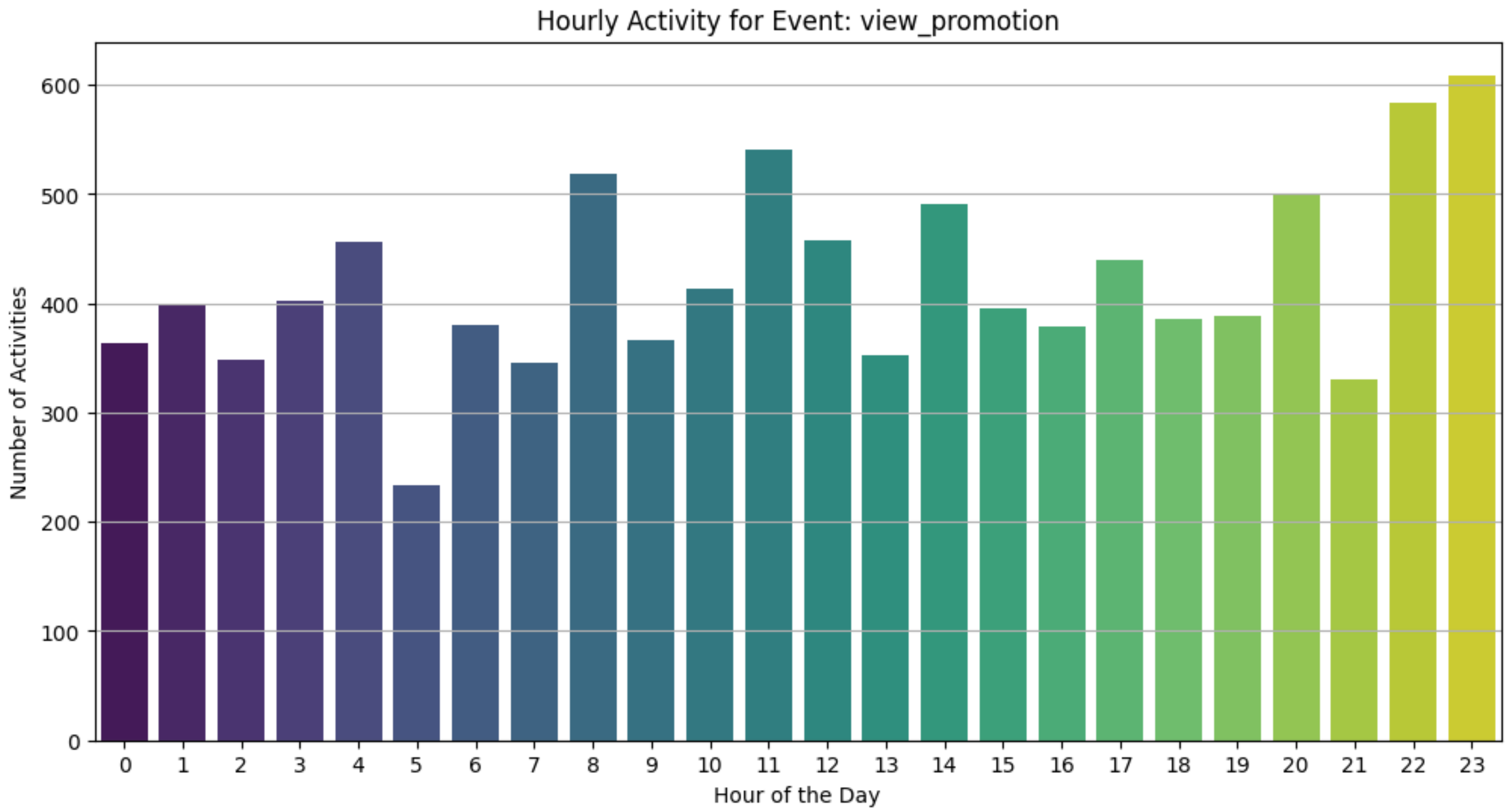
select_item : Looking at the user events in the chart, it is seen that the Select Item event conducts mostly between 6:00 and 12:00.
view_item : Looking at the user events in the chart, it is seen that the View Item event conducts mostly between 00:00 and 6:00.



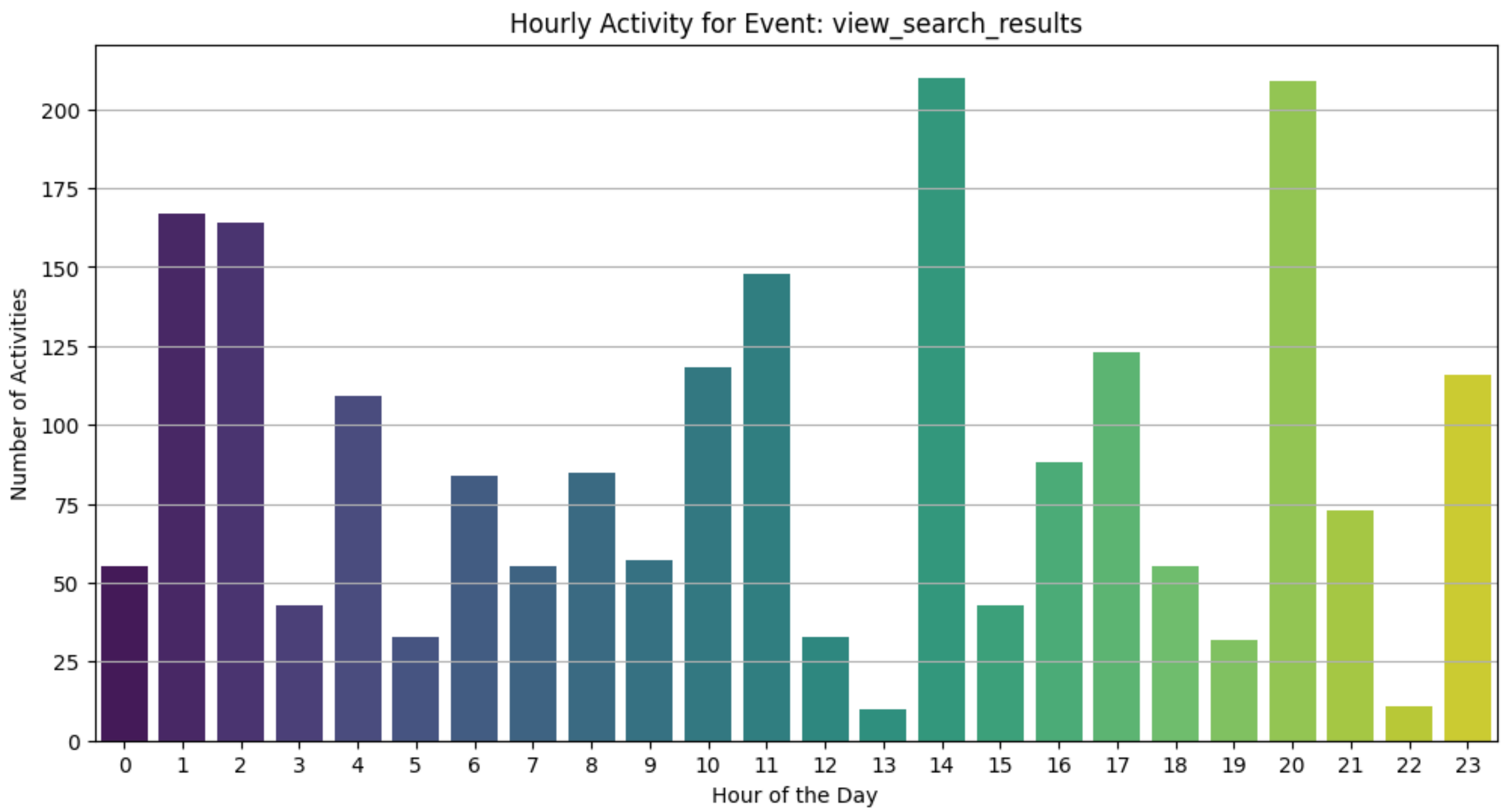


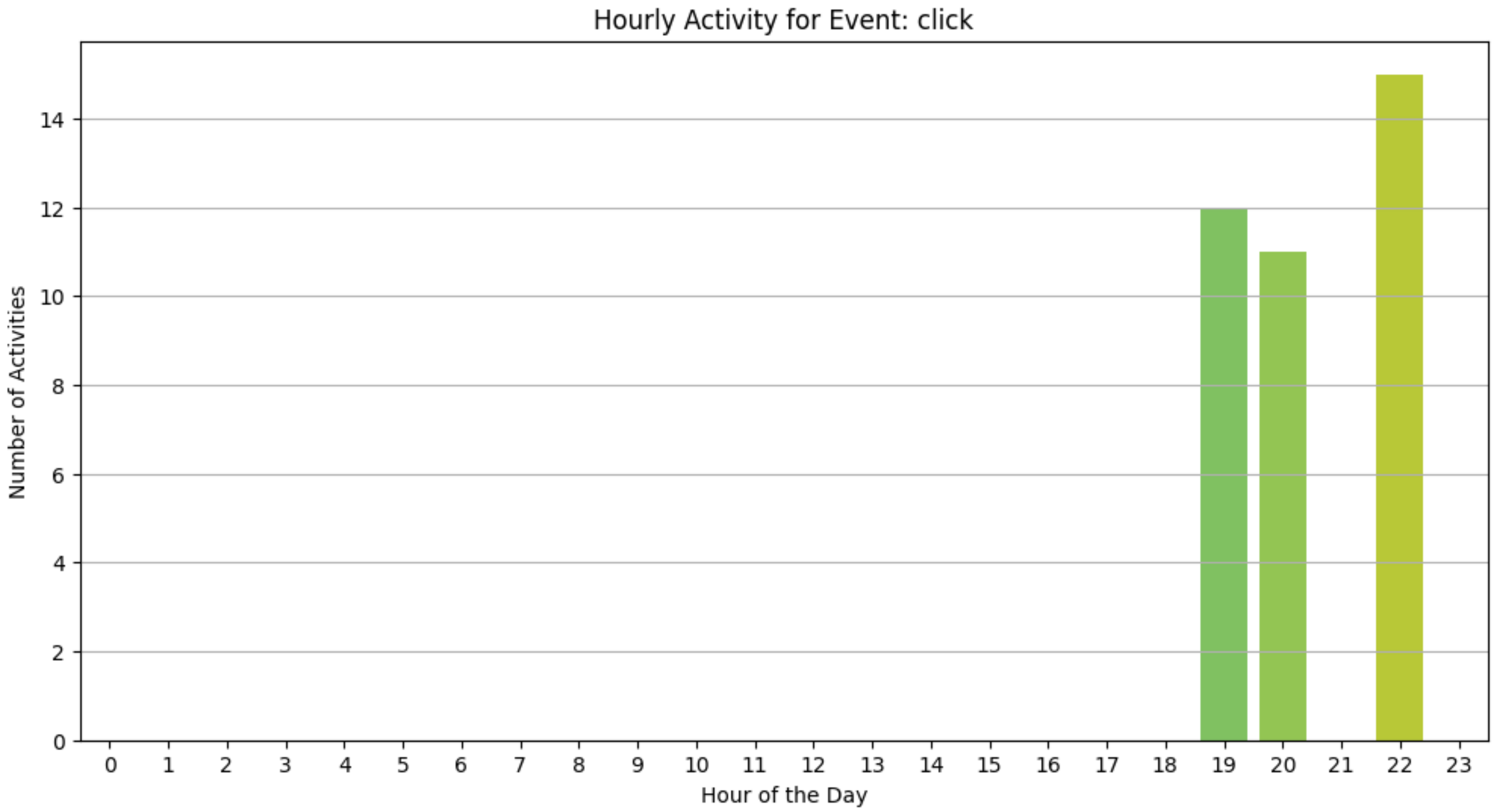
user_engagement : Looking at the user events in the graph, it is seen that the User Engagement event conducts mostly between 8:00 and 14:00.
session_start : Looking at the user events in the graph, it is seen that the Session Start event conducts mostly between 9:00 and 15:00.





select_promotion : Looking at the user events in the chart, it is seen that the Select Promotion event conducts mostly between 2:00 and 8:00.
view_promotion : Looking at the user events in the chart, it is seen that the View Promotion event conducts mostly between 18:00 and 00:00.





click : Looking at the user events in the chart, it is seen that the Click event conducts mostly between 18:00 and 00:00.
view_search_results : Looking at the user events in the graph, it is seen that the View Search Results event conducts mostly between 14:00 and 20:00.

User Purchase Analysis

When we analyze the graphs, we can see the stages of the user entering the website and shopping hourly. Users mostly enter the website in the evening hours and add payment and shipping information. Afterwards, they continue with the promotion selection and purchase steps. However, when we examine the purchase graph, we see that there is a decrease in the purchase phase of users.

Promotion Selection Analysis

When we examine the graphs, users' promotion views are high. However, when we look at the promotion selection, we see that there is a decrease in user transactions. The reason for this is that when the right promotions are not given to users, we might say that users do not shop or use any promotions. In addition, improvements may need to be made in the marketing strategies made here. In this way, we can improve users' use of promotions and purchases.

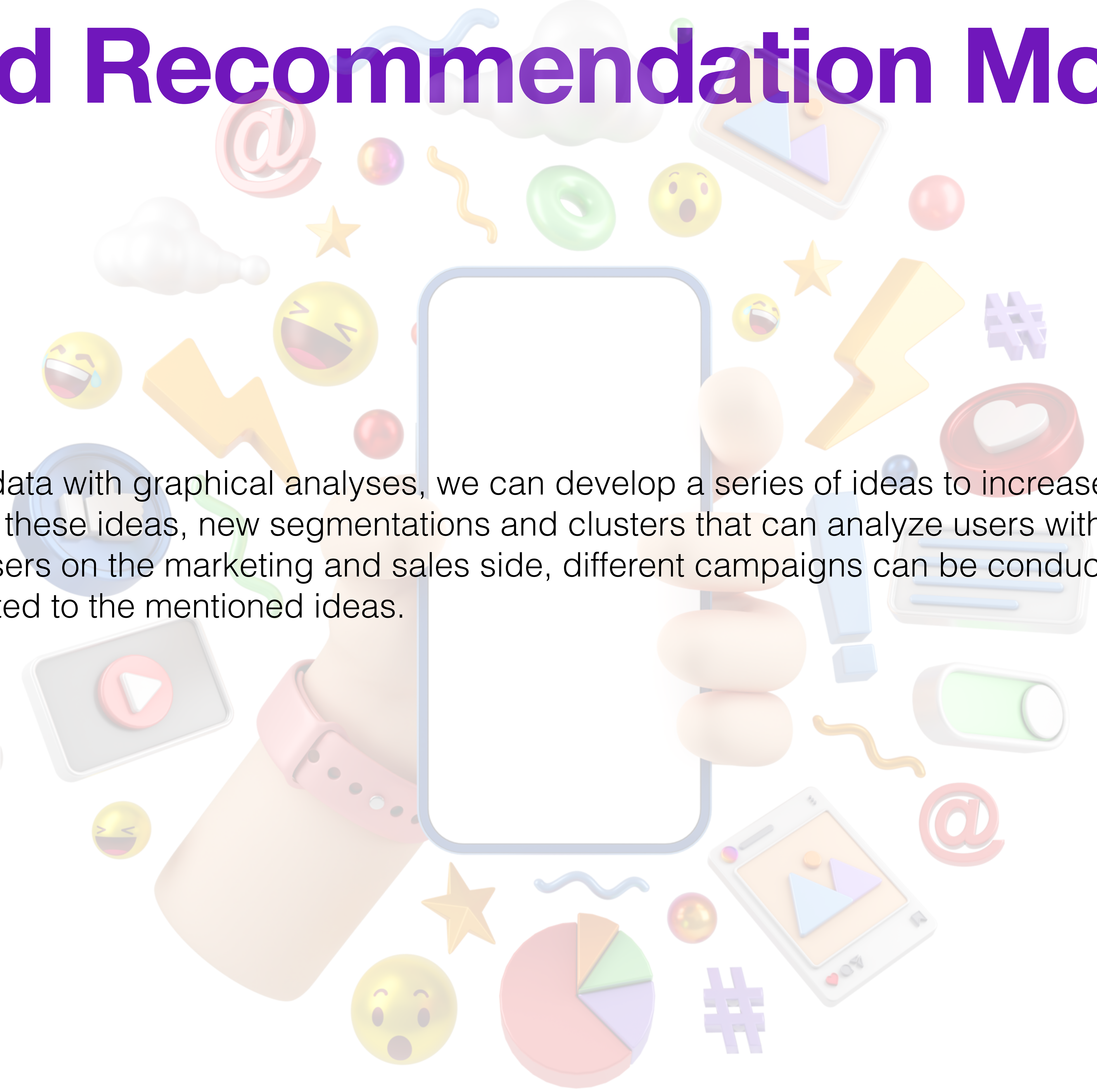
User Product Purchase

When we examine the graphs, we understand that users examine the products more, but they do not select the products and add them to the cart. The right marketing strategies may not have been applied here. Therefore, it causes a decrease in the user's purchase of the product. Here, in the same way, in order to increase sales, we can recommend both improving promotions and improving the product detail pages as a user interface.

Insight And Recommendation

Insight And Recommendation Models

When we examine user activity data with graphical analyses, we can develop a series of ideas to increase sales and improve our marketing strategies. Along with these ideas, new segmentations and clusters that can analyze users within the website can be extracted. Thus, by clustering users on the marketing and sales side, different campaigns can be conducted. I will add a few model development graphs related to the mentioned ideas.



User Segmentation: We can create various user segments based on features such as users' geographic locations, the types of devices they use, and their operating systems. This segmentation can increase engagement by organizing marketing campaigns specific to each user group and personalizing the user experience. For example, we can offer content optimized for mobile device users or highlight local events for users in a specific geographic region.

Activity Forecasting: We can develop a model to predict activities users might perform (such as 'page_view' or 'add_to_cart'). This model allows us to better understand the actions preferred by users and to predict their future movements. Such a model would enable us to determine which content users are more interested in and to offer them more relevant content.

Traffic Source Analysis: Different traffic sources (organic search, social media, direct traffic, etc.) have various effects on user engagement patterns. By analyzing the impact of each source on user behavior, we can more effectively direct our marketing campaigns and budget allocation. For instance, we can determine which platforms are most efficient for boosting the effectiveness of social media campaigns and manage our advertising budget more effectively in light of this information.

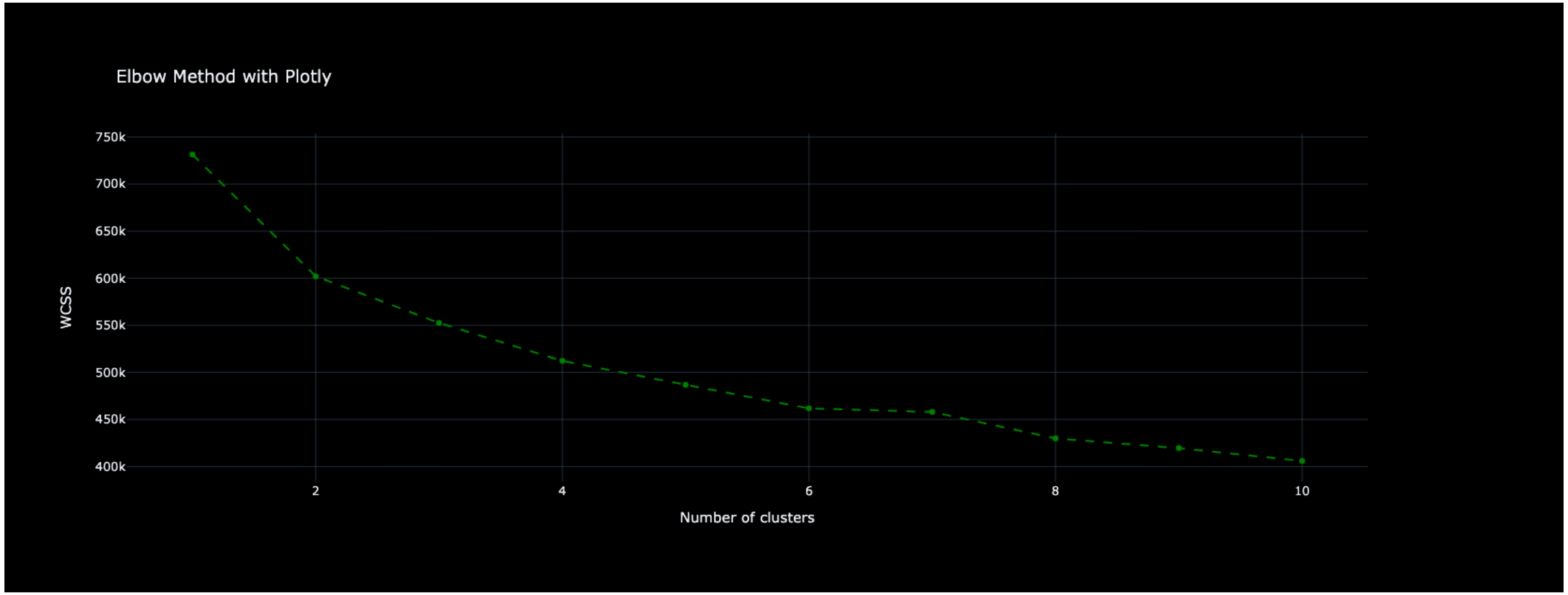
LTV (Lifetime Value) Prediction: A model that predicts the lifetime value (LTV) of users shows us which customer segments have the highest revenue potential. Such a model allows us to customize our marketing and customer relationship management strategies according to specific user groups. For example, we can develop special loyalty programs or incentives targeted at users with high LTV to increase customer loyalty and revenues in the long term.

Time Series Analysis: By using event timestamps, you can analyze trends and patterns over time in website traffic and activities. This can help you develop strategies to increase traffic at certain hours or days, for example.

Anomaly Detection: A model that detects unexpected user behaviors or activities can be used to identify security breaches or system errors at an early stage.

User Segmentation Model

The dataset for user segmentation analysis has been converted into a numerical format using one-hot encoding, making it ready for the modeling process. Key variables considered in the analysis include country, region, device category, and operating system. The k-means clustering analysis, performed using the Elbow method, has been utilized to determine the optimal number of clusters, with the analysis of WCSS values serving as the basis for this determination.
The graph examining WCSS values reveals a 'elbow' point, indicating the decreasing total internal variance with the increase in the number of clusters, and suggesting the optimal number of clusters. This analysis enables effective segmentation of users based on behavioral characteristics, thereby allowing for more informed execution of marketing strategies and campaigns. The modeling process provides valuable insights for a deep understanding of customer behaviors and for the personalization of marketing efforts.



Traffic Source Analysis Model

In this study, a machine learning model was used to classify the effects of traffic sources (organic, social media, etc.) on user behavior. With an overall accuracy of 81%, the model was successful in effectively classifying traffic sources. In particular, the 'organic' and '<Other>' categories stood out as classes that the model recognized well with high precision, recall and F1-scores. In contrast, the '(data deleted)' category performed less well, suggesting that the model struggled to correctly identify and classify data belonging to this category.

These results provide valuable insights for more effective planning of marketing strategies and budget allocation. User behaviors from traffic sources where the model performs well provide important data for optimizing marketing activities for these sources. On the other hand, the low-performing classes show that the model needs to be further improved and that more in-depth analysis of the data of these classes is required.

In conclusion, this model contributes to the development of marketing strategies and a better understanding of traffic sources, but further work is needed to improve the precision and coverage of the model. Feature engineering and hyperparameter tuning, with a particular focus on underperforming classes, will improve the overall performance of the model.