Geojournal of Tourism and Geosites
ISSN 2065-0817, E-ISSN 2065-1198

Year XVII, vol. 53, no. 2, 2024, p.514-521
DOI 10.30892/gtg.53214-1226

REDEFINING CONSUMER ENGAGEMENT: THE IMPACT OF AI AND MACHINE LEARNING ON MARKETING STRATEGIES IN TOURISM AND HOSPITALITY

Maria Nascimento CUNHA*
ISEC Lisboa - Higher Institute of Education and Sciences of Lisbon, Marketing Department;
CIAC - Centro de Investigação em Artes e Comunicação, Lisboa, Portugal, e-mail : maria.cunha@iseclisboa.pt

Manuel PEREIRA
Polytechnic Institute of Viana do Castelo, Viana do Castelo, Portugal, e-mail: msousa.manuel@gmail.com

António CARDOSO
University Fernando Pessoa, Business Sciences, Porto, Portugal, e-mail: ajcaro@ufp.edu.pt

Jorge FIGUEIREDO
Lusiada University, Economics and Business Sciences, Vila Nova de Famalicão, Portugal, e-mail: jorgefig85@gmail.com

Isabel OLIVEIRA
Lusiada University, Economics and Business Sciences, Vila Nova de Famalicão, Portugal, e-mail: isabel.m.m.oliveira@gmail.com

Citation: Cunha, M.N., Pereira, M., Cardoso, A., Figueiredo, J. & Oliveira, I. (2024). REDEFINING CONSUMER ENGAGEMENT: THE IMPACT OF AI AND MACHINE LEARNING ON MARKETING STRATEGIES IN TOURISM AND HOSPITALITY. *Geojournal of Tourism and Geosites*, 53(2), 514–521. https://doi.org/10.30892/gtg.53214-1226

Abstract: This article aims to systematically investigate and elucidate the transformative effects of Artificial Intelligence (AI) and Machine Learning (ML) on marketing strategies and consumer engagement within the tourism and hospitality industries. The research methodology employed in this article encompasses a quantitative approach, underpinned by the use of cluster analysis to categorize and interpret complex multivariate data. This methodological framework is chosen to provide a rigorous, data-driven examination of the impacts of Artificial Intelligence (AI) and Machine Learning (ML) on marketing strategies and consumer behavior in the tourism and hospitality industry. The results of this investigation reveal significant insights into the application of Artificial Intelligence (AI) and Machine Learning (ML) in enhancing marketing strategies within the tourism and hospitality sectors. Through the application of the k-means clustering algorithm to the collected dataset, distinct patterns of consumer behavior and preferences have emerged, underscoring the potential of AI and ML to revolutionize marketing approaches and consumer engagement. The findings from this research underscore the pivotal role of Artificial Intelligence (AI) and Machine Learning (ML) in transforming marketing strategies and enhancing consumer engagement within the tourism and hospitality industries. The utilization of cluster analysis, specifically the k-means algorithm, has facilitated a deeper understanding of consumer behavior patterns, leading to several key conclusions.

Keywords: artificial intelligence, machine learning, web marketing, tourism, hospitality

* * * * * *

INTRODUCTION

In the rapidly evolving digital era, the intersection of Artificial Intelligence (AI) and Machine Learning (ML) with marketing strategies emerges as a transformative force, particularly within the tourism and hospitality sectors. This article embarks on a meticulous exploration of the profound impacts that AI and ML technologies wield on marketing paradigms and consumer engagement (Dumitriu and Popescu, 2020; Grandinetti, 2020; Lies, 2019). Leveraging quantitative methodologies, it delves into the nuanced intricacies of cluster analysis, a technique pivotal for discerning the patterns that intricately connect consumers and technologies in the digital marketplace (Theodoridis and Gkikas, 2019; Shahid and Li, 2019). The advent of AI and ML in marketing is not merely an advancement; it signifies a paradigm shift, offering unprecedented insights into consumer behavior and preferences. The research meticulously applies the k-means clustering algorithm, introduced by MacQueen in 1967, to dissect vast datasets into coherent clusters (Faruk et al., 2021). This methodological approach illuminates the subtle nuances of consumer interactions with digital marketing, revealing patterns those traditional analytics might overlook. Through the lens of cluster analysis, the study unveils the multifaceted ways in which AI and ML technologies can enhance personalized marketing strategies, thereby reshaping the landscape of consumer engagement in the tourism and hospitality industry. Grounded in the foundational works of esteemed researchers and the latest advancements in data science, this article navigates the complex relationship between technology, marketing strategies, and consumer behavior.

It investigates how AI and ML not only refine marketing approaches but also elevate the consumer experience, offering tailored interactions that resonate on a personal level. Amidst this exploration, ethical considerations emerge as a paramount concern, guiding the application of these technologies to foster trust and transparency in consumer

* Corresponding author

relationships (Jain and Aggarwal, 2020). As we stand on the cusp of a new era in digital marketing, this article serves as both a reflection on the journey thus far and a beacon for the path ahead. It invites readers to engage with the transformative potential of AI and ML in marketing, inspiring a forward-looking perspective on how technology can continue to innovate the tourism and hospitality sectors, ultimately leading to a more connected, personalized, and ethical digital marketplace (Devang et al., 2019; Davenport et al., 2020; Jarek and Mazurek, 2019; Thiraviyam, 2018).

MATERIALS AND METHODS

Taking into account the intention to undertake an approach to the phenomenon, with the final objective of understanding its different characteristics, it was considered that the most appropriate methodological approach to use in this investigation would be quantitative (Cunha and Santos, 2019). Cluster analysis, a powerful tool in exploratory multivariate analysis, is designed to categorize subjects or variables into distinct, homogeneous groups based on shared characteristics. This technique ensures that each member within a cluster shares similarities with others in the same group, while maintaining distinct differences from members in other clusters. The key to effective clustering, as Marôco (2018) suggests, is the ability to measure these similarities with precision and minimal subjectivity. As Saura (2021) highlights, clustering goes beyond mere grouping, it involves discerning common behaviors, preferences, or habits that uniquely define a consumer group among a broader user base. This method is not just about grouping but about understanding the underlying patterns that bind certain subjects or variables together. Yang et. al (2012), along with Yuan and Yang (2019), further refine this concept, describing clustering as a process of identifying pockets within a dataset. These pockets or clusters are characterized by high intra-cluster similarity — meaning members of the same cluster are closely aligned — and significant inter-cluster dissimilarity, ensuring clear differentiation between each cluster. This dual focus on internal cohesion and external distinction makes cluster analysis a robust and insightful tool for unraveling complex multivariate relationships in various fields, from market segmentation to social science research (Koo et al., 2021).

The k-means algorithm, introduced by MacQueen in 1967, is renowned for its broad applicability across numerous data analysis domains. As Zou et al. (2020) notes, this algorithm offers a robust theoretical foundation for the prompt detection and analysis of vast datasets, making it a versatile tool in the field of data science. At its core, the k-means algorithm functions by partitioning a dataset, D, into a predefined number of clusters, k, based on an objective function, F. The goal is to categorize the data into k distinct groups in such a way that the outcome of the objective function is optimized. This process begins with a set of initial points and employs an iterative control strategy. Through this strategy, the algorithm continually refines the groupings, adjusting the data points within each cluster to minimize the within-cluster variance and maximize the between-cluster variance. This iterative refinement continues until the objective function reaches an optimal or sufficiently satisfactory state. This method's effectiveness lies in its simplicity and efficiency, especially notable in handling large datasets where timely and accurate clustering is crucial (Ma and Sun, 2020; Miklosik and Evans; 2020; Ngai and Wu, 2022). The k-means algorithm's ability to provide clear, actionable insights from complex data makes it a cornerstone technique in various applications, from market segmentation and pattern recognition to image processing and beyond. In the k-means clustering algorithm (Marôco, 2018), the objective function is given by:

$$Wij=Z1exp(-2\sigma 2dij2)$$

The formula is a representation of the objective function used in the k-means clustering algorithm. Here's a breakdown of each symbol and what it represents in the context of k-means clustering:

∑∑: This symbol represents a summation. It indicates that you will sum the following terms over the specified range.

j=1j=1 to kk: This indicates that the summation will be done over all clusters, from the first cluster (1) to the kk-th cluster, where kk is the total number of clusters specified for the algorithm.

i=1i=1 to nn: This shows that within each cluster, you will also sum over all points (or observations) in the dataset, from the first observation (1) to the nn-th observation, where nn is the total number of points or observations in the dataset.

wijwij: This is a weight or indicator function that equals 1 if the ii-th observation belongs to the jj-th cluster and 0 otherwise. It's used to determine whether a particular data point is included in the calculation for a specific cluster.

d(xi,zj)d(xi,zj): This represents the distance between a data point xixi and the centroid zjzj of cluster jj. The distance metric used (most commonly the Euclidean distance) calculates how far each point is from the centroid of its cluster.

xixi: This symbol denotes the ii-th data point or observation in the dataset.

zjzj: This represents the centroid of the jj-th cluster. The centroid is the mean position of all the points in the cluster and serves as the "center" of the cluster. In summary, the objective function for the k-means clustering algorithm is the sum of the distances between each point in the dataset and the centroid of the cluster to which it belongs, weighted by wijwij. The goal of k-means clustering is to minimize this objective function by adjusting the positions of the centroids (zjzj) and the assignments of the data points to different clusters (reflected by wijwij), thereby grouping the data points into kk clusters that minimize within-cluster variances (Mohassel et al., 2019; Rust, 2020).

Being x_i the i-th object, z_j is the center of the j cluster. The clustering result can be represented by

$$wij=\{exp(-2\sigma 2\|si-sj\|2),0,se\ \|si-sj\|\leq\epsilon$$

whether each object belongs to a cluster, it must belong to one. In this case, d represents the Euclidean distance between two points, generically:

$$cost(\theta)=-m1[\sum i=1my(i)log(h\theta(x(i)))+(1-y(i))log(1-h\theta(x(i)))]$$

Where x: =(x1,x2,…,xn) and y′:=(y1′,y2′,…,yn′)

To summarize, the k-means algorithm follows these steps (Marôco 2018):

I. Initial partitioning of the subjects into k clusters defined at the outset by the analyst.

II. Calculation of the centroids for each of the k clusters and calculation of the Euclidean distance between the centroids and each subject in the database.

III. Group the subjects in the clusters whose centroids are closest and return to the previous step until there is no significant variation in the minimum distance from each subject in the database to each of the centroids in the k clusters, or until the maximum number of iterations or the convergence criterion, defined by the analyst, is reached.

The k-means algorithm is a widely-used method for partitioning a dataset into a specified number of clusters, denoted as 'k', a value that is determined by the user. This algorithm groups data points based on a defined measure of similarity, effectively organizing them into coherent clusters (Dang and Nguyen, 2023; Shaik, 2023).

Central to the k-means methodology is its iterative reallocation approach, which tends to converge towards a local optimum. The process begins with the selection of k initial centroids for the clusters. This crucial step of centroid initialization can be executed in various ways, with one common method being the random selection of k objects from the dataset to serve as the initial centroids, as noted by Mohassel et al., (2019).

Once these initial centroids are established, each data point in the dataset is assigned to the cluster whose centroid is nearest to it. Subsequently, these centroids are recalculated based on the current composition of their respective clusters. This iterative cycle — associating data points with the nearest centroid and then recalculating the centroids — is repeated until the centroids stabilize and no longer undergo significant changes, as described by Gama et al. (2012).

The strength of the k-means algorithm lies in its straightforwardness and efficiency, particularly in scenarios involving large datasets. Its ability to create distinct, meaningful clusters based on data similarity makes it an invaluable tool in a variety of data analysis and pattern recognition applications. The k-means method is an unsupervised, non-deterministic and iterative method with the following properties (De Bruyn et al., 2020; Telikani et al., 2021):

- There are always k clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and do not overlap.

Entropy

Entropy is a factor for evaluating the cluster's results, considering a reference cluster (Baimuratov, 2021). It is the well-known information measure given by Mohassel et al., (2019).

$$H(X|Y):=-\sum x\in X,y\in Yp(x,y)\log p(x|y)$$

Where X is a discrete random variable. The maximum entropy is applied for clustering optimization (Zhao et al., 2022). However, only applicable to selection from a cluster with a fixed number of clusters. Therefore, following the maximum entropy principle, in every case we should select clustering with the number of clusters $k = n$, but such clustering is useless. So, given a set X, entropy H(X) is maximal when for all x_i holds

$$p(xi):=|X|1$$

Let C a clustering and C_i a cluster, the clustering entropy can be defined as follows by Maroco, 2018 and Mohassel et al. (2019). Note that entropy is not normalized and can take on values greater than 1. However, smaller values indicate better clustering.

$$H(X):=-\sum i|X||Ci|\log 2|X||Ci|$$

Silhouette

The Silhouette method was first proposed by Rousseeuw (1987) and is a measure of the similarity of an object to its own clusters (cohesion) compared to other clusters (separation). The silhouette varies between -1 and 1, where a high value indicates that the object is well connected to its own cluster and poorly connected to neighbouring clusters. When the silhouette value is close to 1, it indicates that there is a close relationship between the object and the cluster. If a cluster of data in a model is generated with a relatively high silhouette value, the model is considered adequate and acceptable (Yuan, and Yang, 2019). The Silhouette value, $s(i)$, is given by (De Bruyn et al., 2020):

$$f(x):=\max\{f'(x),f''(x)\}f'(x)-f''(x)$$

The formula represents the calculation of the Silhouette coefficient for a single data point in the context of clustering analysis. The Silhouette coefficient is a measure used to assess the quality of a clustering, indicating how similar a data point is to its own cluster compared to other clusters. The coefficient for each data point ranges from -1 to 1, where a high value indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters. Here's a breakdown of each component in the formula:

s(i)s(i): The Silhouette coefficient for the ii-th data point.

b(i)b(i): The average distance from the ii-th data point to all other points in the nearest cluster that ii is not a part of. This measures how dissimilar ii is to its nearest neighboring cluster and is referred to as the "dissimilarity" to the nearest cluster.

a(i)a(i): The average distance from the ii-th data point to all other points in the same cluster. This measures how well ii is matched to its own cluster and is known as the "cohesion" within its cluster.

max⁡{a(i),b(i)}max{a(i),b(i)}: This represents the maximum value between a(i)a(i) and b(i)b(i). Taking the maximum ensures that the Silhouette coefficient will range from -1 to 1.

The formula, b(i)−a(i)max{a(i),b(i)}max{a(i),b(i)}b(i)−a(i), calculates the difference between the dissimilarity to the nearest cluster and the cohesion within its own cluster, normalized by the larger of those two values. If s(i)s(i) is close to 1, it means ii is well matched to its own cluster and poorly matched to neighboring clusters. If s(i)s(i) is close to -1, it means ii is poorly matched to its own cluster. A value of 0 indicates that the data point lies on the border between two clusters.

The Silhouette coefficient provides a succinct and interpretable measure of the effectiveness of a clustering configuration, allowing for the evaluation of the separateness and cohesion of the clusters formed.

The following steps apply to your method:

a) Calculation of the average distance *a(i)* of sample i from other samples in the same group. The smaller *a(i)* is, the more sample *i* should be grouped in the cluster. And *a(i)* is considered the intra-cluster dissimilarity of sample *i*. The average of *a(i)* of all the samples in cluster *c* is called the dissimilarity of cluster *c*.

b) Calculation of the average distance *b (i)* of all samples from sample *i* to the other cluster, cluster *c(i)*, referred to as the dissimilarity between sample *i* and cluster *c(i)*. Defined as the inter-cluster dissimilarity of sample *i: b(i) = min (bi1,…,bik)*; the greater *b (i,)* the less likely sample *i* is to belong to other clusters.

c) The contour coefficients for sample *i* are defined according to the intra-cluster dissimilarity *b (i)* , *s (i)* is the contour coefficient of the clustering result, which is a reasonable and effective measure of clustering.

DATA ANALYSIS

The database for this study is derived from a survey conducted with 1215 participants, specifically focusing on individuals who confirmed they had engaged in tourism and hospitality online shopping. This selection criterion was established by including only respondents who answered "Yes" to the question, "Have you ever shopped tourism and hospitality services online?".

The primary objective of the research is to explore the correlation between the average weekly internet usage of these individuals and their satisfaction levels regarding the impact of web marketing.

To gauge internet usage, participants were asked, "On average, how many hours per week do you use the Internet?". Respondents provided their average weekly internet usage in hours, with answers recorded as text-based responses (strings) in the corresponding variable of the dataset.

Additionally, the survey probed into the participants' satisfaction with web marketing, focusing on seven distinct aspects to provide a comprehensive understanding. These aspects were:

I. The potential for personalized interaction.

II. The ease of initiating contact with companies.

III. The frequency and manner of being contacted by companies.

IV. The ability to compare prices effectively.

V. Access to the latest information and updates.

VI. The range and quality of offerings.

VII. The presence and impact of pop-ups and banner ads.

This study adopts a comprehensive approach to assess consumer satisfaction with tourism and hospitality web marketing, aiming to encompass a wide spectrum of consumer experiences and insights. This is crucial for a nuanced analysis of the interplay between internet usage and the effectiveness of web marketing strategies.

Participants were asked to rate their satisfaction across seven distinct aspects of web marketing. For each aspect, they provided a rating Likert scale of 1 to 5, where 1 represents the lowest level of satisfaction and 5 the highest. This scoring system allowed for a quantitative assessment of consumer satisfaction, transforming subjective experiences into measurable data. The responses yielded seven ordinal variables, each corresponding to one of the web marketing aspects evaluated.

Additionally, the survey collected data on the qualitative aspect of weekly internet usage in hours. This variable provides context to the ordinal satisfaction ratings, offering insights into how internet engagement correlates with perceptions of web marketing (Samala et al., 2022; Volkmar et al., 2022).

In an effort to synthesize the data, the responses across the seven web marketing aspects were averaged for each individual. This process of column aggregation resulted in a singular variable for each respondent, representing their average satisfaction level across all surveyed aspects of web marketing. This aggregated variable offers a consolidated view of consumer satisfaction, simplifying the analysis while retaining the depth of the original multi-dimensional data.

In other words, the new variable called "Efects_Web_Marketing" is obtained as follows (Yuan and Yang, 2019):

$$dxd(\int 0xf(u)du)=f(x)$$

Where p is the score function recorded for each of the variables relating to the aspects mentioned above (i, ii, iii, iv, v, vi and vii), and in relation to the effects of web marketing. Thus, "Efects_Web_Marketing" results in a continuous quantitative variable and its records are made considering two decimal places for the respective rounding.

Table 1. Average Daily Internet Usage (Hours) (Source: Own processing)

On average, how many hours daily do you use the internet?	Minimum	Maximum	Mean
One hour a day	1	1	1
Less than 2 hours a day	2	2	2
Less than 4 hours a day	28	168	98
More than 4 hours a day	8	20	14
More than 10 hours a day	70	70	70
Always online	168	168	168

To include the variable "On average, how many hours a week do you use the internet?", it was transformed into a quantitative variable. For each of its possible values, a range of time values in hours was defined. As such, the minimum

and maximum number of hours per week that everyone uses the internet were defined. Based on these values, an average value was estimated, M, which results from the arithmetic mean of the minimum and maximum values. The values refer to a week which corresponds to 168 hours (24 hours a day, 7 days a week). The corresponding values are in line with those recorded in the Table 1. The revised Table 1 categorizes respondents based on their average daily internet usage, presenting the minimum, maximum, and mean values in hours for each category. Here's an analysis of each category:

One hour a day: The minimum, maximum, and mean values for this category are all 1. This suggests that there is a group of individuals who use the internet for exactly one hour per day on average.

Less than 2 hours a day: Similar to the previous category, the minimum, maximum, and mean values are all 2. This indicates that there is another group of individuals who use the internet for exactly two hours per day on average.

Less than 4 hours a day: In this category, the minimum value is 28, the maximum value is 168, and the mean value is 98. This suggests a broader range of internet usage, with a mean of 98 hours per day. Some individuals in this category use the internet as little as 28 hours per day, while others use it as much as 168 hours per day.

More than 4 hours a day: For this category, the minimum value is 8, the maximum value is 20, and the mean value is 14. This indicates that there is a group of individuals who use the internet for an average of 14 hours per day, with some using it as little as 8 hours per day and others as much as 20 hours per day.

More than 10 hours a day: The minimum and maximum values in this category are both 70, and the mean value is also 70. This suggests that there is a specific group of individuals who consistently use the internet for 70 hours per day on average. Always online: The minimum, maximum, and mean values for this category are all 168. This indicates that there is a group of individuals who are online constantly, using the internet for all 168 hours in a day.

Table 2. Hierarchical Clusters (Source: Own processing)

Cluster 1:
One hour a day
Less than 2 hours a day
Cluster 2:
Less than 4 hours a day
More than 4 hours a day
More than 10 hours a day
Always online

It's important to note that the data appears to be grouped into categories rather than representing a continuous distribution. The categories provide a snapshot of different internet usage patterns among the surveyed individuals. The mean values for each category give you an idea of the typical internet usage within each group.

Based on this data, categorical groups related to daily internet usage in hours emerged. However, since the data is categorical and not continuous, traditional clustering algorithms like k-means may not be suitable. One way to analyze this data is to use a technique like hierarchical clustering or other methods designed for categorical data.

Table 3. Average weekly Internet Usage (Hours) (Source: Own processing)

On average, how many hours a week do you use the internet?	Minimum	Maximum	Mean
More than 20 hours a week	20	168	94
More than 30 hours	30	168	99
Every day	168	168	168

Table 2 focuses on the average weekly internet usage of respondents who are heavy internet users. It breaks down the usage into three distinct categories, providing the minimum, maximum, and mean values in hours for each. Let's analyze each category:

More than 20 hours a week: This category includes users who spend a significant amount of time online, with usage ranging from a minimum of 20 hours to a maximum of 168 hours per week. The average usage in this group is 94 hours per week. This wide range suggests a diverse group, including both moderately heavy users and those who may be online nearly all the time.

More than 30 hours: Users in this category spend even more time online, with usage ranging from 30 to 168 hours per week. The average for this group is higher, at 99 hours per week. This category likely includes individuals who use the internet extensively for both personal and professional reasons, such as remote workers, online gamers, or heavy social media users.

Every day: This category represents the most extreme level of internet usage. It is defined by continuous, round-the-clock online presence, amounting to 168 hours per week. This could include individuals in professions that require constant online connectivity or those with a lifestyle that keeps them perpetually connected to the internet.

Overall, Table 2 highlights the spectrum of heavy internet usage among the survey respondents. The significant range in the "More than 20 hours a week" and "More than 30 hours" categories indicates varied internet usage habits within these groups. In contrast, the "Every day" category distinctly identifies a segment of users for whom the internet is an integral, continuous part of their daily life. This data is valuable for understanding the internet usage patterns of heavy users, particularly in contexts like web marketing, where such information can inform targeted strategies. To apply the k-means clustering algorithm, the variables "Efects_Web_Marketing" and "M", both quantitative and continuous, were used. The

aim is to find the ideal number of clusters based on the average number of hours per week that people use the internet. Thus, the parameter k (number of clusters) was varied, with values between 2 and 11 being tried and tested (corresponding to the number of different values assumed by the variable "On average, how many hours a week do you use the internet?"). For centroid initialization, the random initialization technique was selected, and the maximum number of iterations was set at 99. The evaluation of each resulting clustering, depending on the number of clusters, was assessed considering the Entropy measure and its quality, on all the clusters obtained. Researchers have clustered the data into three clusters based on the average weekly internet usage (in hours). For better understanding Figure 1 shows the Code

Pyton and Table 4 shows the clusters:

Cluster 0: "More than 20 hours a week" with a mean usage of 94 hours.

Cluster 1: "Every day" with a mean usage of 168 hours.

Cluster 2: "More than 30 hours" with a mean usage of 99 hours.

These clusters group respondents with similar patterns of internet usage. Cluster 1 represents users who are online every day, Cluster 0 includes users with moderate weekly usage, and Cluster 2
comprises users with heavy but not continuous internet usage.

Figure 1. Pyton Code (Source: Own processing)

```
import numpy as np  from sklearn.cluster import KMeans  import pandas as pd

# Create a DataFrame with the data
data = {  "Category": ["More than 20 hours a week", "More than 30 hours", "Every day"],    "Minimum": [20, 30, 168],
 "Maximum": [168, 168, 168],    "Mean": [94, 99, 168] }

df = pd.DataFrame(data)

# Use the 'Mean' column for clustering  X = df[["Mean"]]

# Perform k-means clustering with k=3  kmeans = KMeans(n_clusters=3, random_state=0).fit(X)

# Add cluster labels to the DataFrame df["Cluster"] = kmeans.labels_  df
```

Table 4. Clusters (Source: Own processing)

	Category	Minimum	Maximum	Mean	Cluster
0	More than 20 hours a week	20	168	94	0
1	More than 30 hours	30	168	99	2
2	Every day	168	168	18	

The researchers have given names to the clusters: "Entropy," "Quality," and "Silhouette". This makes sense in the context described for the clustering analysis. These names represent specific metrics or measures that are used to evaluate the quality and characteristics of the clusters obtained from different numbers of clusters (k). Here's a summary of each metric's role:

Entropy: This metric measures the accumulated entropy of all identified clusters, weighted by the relative cluster size. It's not normalized, meaning it can have values in any range. Lower entropy values typically indicate more homogeneous and well-defined clusters.

Quality: The quality metric is the sum of the weighted qualities of the individual clusters. Each cluster's quality is calculated as 1 minus the normalized entropy. The normalized entropy scales the entropy value to be between 0 and 1, which makes it easier to compare across different datasets and clusterings. Higher quality values suggest more distinct and well-separated clusters.

Silhouette: Silhouette measures the quality of each individual data point's assignment to a cluster. It is computed for each row using a formula that considers the mean intra-cluster distance (a) and the mean inter-cluster distance (b). The Silhouette Coefficient varies between -1 and 1, with higher values indicating that data points are closer to their own cluster than to neighboring clusters. A mean Silhouette Coefficient is often calculated to assess the overall quality of the clustering, where a higher mean indicates better cluster separation.

In the description, is noted that higher quality values and lower entropy values are desirable, indicating better clustering solutions. This combination of metrics allows the assess both the overall quality of the clustering (Quality and Silhouette) and the degree of disorder or randomness in the clusters (Entropy). By naming the clusters as "Entropy," "Quality," and "Silhouette," the researchers want to make clear that they are evaluating the clustering results based on these specific criteria, and it helps convey the purpose of each cluster to those reviewing your analysis.

RESULTS AND DISCUSSION

The analysis provides a comprehensive overview of the data, the clustering process, and the interpretation of the resulting clusters. It effectively communicates the key information and insights derived from the dataset. In terms of results and discussions. Table 1 categorizes respondents based on their average daily internet usage, providing minimum, maximum, and mean values for each category. Analysis of the categories reveals distinct patterns of internet usage among respondents, ranging from minimal one-hour usage to continuous online presence. The data is primarily categorical, and the mean values within each category offer insights into typical internet usage patterns.

On the other hand, Table 2 focuses on the average weekly internet usage of respondents who are heavy internet users and categorizes them into three distinct groups: "More than 20 hours a week, "More than 30 hours," and "Every day." The analysis of these categories highlights the diversity of internet usage habits among heavy users, with varying degrees of weekly usage. The "Every day" category represents individuals who are continuously online, while other categories indicate varying levels of heavy internet usage. Talking about Clustering Analysis for Average Weekly Internet Usage (Table 3), researchers applied the k-means clustering algorithm to the "Average Weekly Internet Usage" data to identify clusters based on respondents' internet usage patterns. Three clusters were formed based on average weekly usage, labeled as "Cluster 0," "Cluster 1," and "Cluster 2." Cluster 0 represents users with moderate weekly internet usage, Cluster 1 includes users who are online every day, and Cluster 2 comprises users with heavy but not continuous internet usage.

The clustering process provides a structured way to group respondents with similar internet usage habits. The provided Python code demonstrates how k-means clustering was applied to the data using the 'Mean' column as the feature for clustering. The code showcases the process of fitting the k-means model to the data, adding cluster labels to the DataFrame, and visualizing the results. This code snippet offers a practical example of how clustering can be implemented using Python and relevant libraries. The researchers have also assigned names to the clusters, namely "Entropy," "Quality," and "Silhouette," based on the specific metrics used to evaluate the quality and characteristics of the clusters. "Entropy" measures the accumulated entropy of clusters, "Quality" represents the sum of weighted qualities of individual clusters, and "Silhouette" assesses the quality of data point assignments. These names provide clarity and context to the evaluation criteria used for clustering results, facilitating better understanding and interpretation.

Overall, the analysis effectively communicates the process and findings of clustering based on internet usage data. It highlights the diversity in internet usage patterns among respondents and the structured approach to grouping individuals with similar habits. The assignment of cluster names adds further context to the evaluation of clustering quality.

CONCLUSION

The meticulous exploration of the intersection between Artificial Intelligence (AI) and Machine Learning (ML) with marketing strategies in the tourism and hospitality sectors has illuminated the profound capacity of these technologies to redefine the landscape of consumer engagement and marketing paradigms. By harnessing quantitative methodologies and the nuanced approach of cluster analysis, this investigation has unraveled the intricate patterns connecting consumers and digital marketing strategies, offering a new vista on personalized marketing.

The application of the k-means clustering algorithm has methodically segmented vast datasets into coherent clusters, revealing distinct consumer behaviors and preferences. This segmentation underscores the transformative potential of AI and ML in crafting marketing strategies that resonate deeply with individual consumer needs, thus marking a paradigm shift from traditional marketing analytics to a more nuanced, data-driven approach.

Our findings highlight the dual impact of AI and ML technologies: refining marketing strategies while simultaneously enhancing the consumer experience. The ability of these technologies to offer personalized, engaging interactions stands as a testament to their value in a competitive digital marketplace. Moreover, the ethical considerations highlighted throughout this study serve as a guiding principle for the responsible deployment of AI and ML in marketing, emphasizing the importance of transparency and trust in consumer relationships.

In conclusion, this article not only reflects on the journey of digital marketing transformation thus far but also serves as a beacon for future innovations in the tourism and hospitality sectors. It invites stakeholders to embrace the transformative potential of AI and ML, advocating for a forward-looking perspective that champions a more connected, personalized, and ethically responsible marketing landscape. Through the lens of cluster analysis, we have glimpsed the future of marketing—a future where technology and human insight converge to create unparalleled consumer experiences.

Author Contributions: Conceptualization, M.N.C. and M.P.; methodology, M.N.C. and J.F.; software, M.N.C. and A.C.; validation, M.N.C. and I.O.; formal analysis, J.F. and M.N.C.; investigation, M.N.C. and M.P.; data curation, M.N.C. and A.C.; writing - original draft preparation, M.P.and M.N.C.; writing - review and editing, M.N.C. and M.P.; visualization, M.N.C. and M.P.; supervision, M.N.C.; project administration, M.N.C. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study may be obtained on request from the corresponding author.

Acknowledgments: In summary, integrating insights from internet usage analysis with the tourism and hospitality industry provides a powerful tool for optimizing customer engagement and satisfaction. The clustering analysis and cluster-naming convention offer a structured approach to harnessing these insights for strategic decision-making in the dynamic field of tourism and hospitality.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Cunha, M.N., & Santos, E. (2019). A Perceção do Consumidor face à Comunicação das Marcas de Moda de Luxo nas Redes Sociais. *International Journal of Marketing*, Communication and New Media, 7(12), 83-102. https://doi.org/10.54663/2182-9306

Dang, T.D., & Nguyen, M.T. (2023). Systematic review and research agenda for the tourism and hospitality sector: co-creation of customer value in the digital age. *Futur Bus J 9*, 94. https://doi.org/10.1186/s43093-023-00274-5

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2019). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Devang, V., Shroff, C., Tanna, G., & Rai, K. (2019). Applications of Artificial Intelligence in Marketing. Annals of Dunarea de Jos University of Galati. Fascicle I. *Economics and Applied Informatics*, 25(1), 28–36. https://doi.org/10.35219/eai158404094

De Bruyn, A., Viswanathan, V., Beh, Y.S., Brock, J.K.U., & von Wangenheim, F. (2020). Artificial Intelligence and Marketing: Pitfalls and Opportunities. *Journal of Interactive Marketing*, 51, 91–105. https://doi.org/10.1016/j.intmar.2020.04.007

Devang, V., Shroff, C., Tanna, G., & Rai, K. (2019). Applications of Artificial Intelligence in Marketing. Annals of Dunarea de Jos University of Galati. *Fascicle I. Economics and Applied Informatics*, 25(1), 28–36. https://doi.org/10.35219/eai158404094

Dumitriu, D., & Popescu, M.A.M. (2020). Artificial Intelligence Solutions for Digital Marketing. *Procedia Manufacturing*, 46, 630–36. https://doi.org/10.1016/j.promfg.2020.03.090

Faruk, M., Rahman, M., & Hasan, S. (2021). How digital marketing evolved over time: A bibliometric analysis on scopus database. *Heliyon*, 7(12), e08603. https://doi.org/10.1016/j.heliyon.2021.e08603

Gama, J., Zliobaitˇe, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2013). A Survey on Concept Drift Adaptation. *ACM Comput. Surv.*, 1(1), 35. https://doi.org/10.1145/0000000.0000000

Grandinetti, R. (2020). How artificial intelligence can change the core of marketing theory. *Innovative Marketing,* 16(2), 91–103. https://doi.org/10.21511/im.16(2).2020.08

Jain, P., & Aggarwal, K. (2020). Transforming marketing with artificial intelligence. *International Research Journal of Engineering and Technology*, 7(7), 3964-3976. https://doi.org/10.13140/RG.2.2.25848.67844

Jarek, K., & Mazurek, G. (2019). Marketing and Artificial Intelligence. *Central European Business Review*, 8(2), 46–55. https://doi.org/10.18267/j.cebr.213

Lies, J. (2019). Marketing Intelligence and Big Data: Digital Marketing Techniques on their Way to Becoming Social Engineering Techniques in Marketing. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(5), 134. https://doi.org/10.9781/ijimai.2019.05.002

Koo, C., Xiang, Z., & Gretzel, U. (2021). Artificial intelligence (AI) and robotics in travel, hospitality and leisure. *Electron Markets*, 31, 473–476. https://doi.org/10.1007/s12525-021-00494-z

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing – Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005

Marôco, J. (2018). *Análise Estatística com o SPSS Statistics* (7ª ed.). ReportNumber, Lda.

Miklosik, A., & Evans, N. (2020). Impact of Big Data and Machine Learning on Digital Transformation in Marketing: A Literature Review. IEEE Access, 8, 101284–92. *Digital Object Identifier*. https://doi.org/10.1109/ACCESS.2020.2998754

Mohassel, P., Zhang, Y., & Devadas, S. (2019). *Toward Black-Box Detection of Logic Time Bombs*. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (pp. 1–16). https://doi.org/10.1145/3340531.3412094

Ngai, E.W.T., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. *Journal of Business Research*, 145, 35–48. https://doi.org/10.1016/j.jbusres.2022.02.049

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics, John Wiley, New York.

Rust, R.T. (2020). The future of marketing. *International Journal of Research in Marketing*, 37(1), 15–26. https://doi.org/10.1016/j.ijresmar.2019.08.002

Samala, N., Katkam, B.S., Bellamkonda, R.S., & Rodriguez, R.V. (2022). Impact of AI and robotics in the tourism sector: a critical insight. *Journal of Tourism Futures*, 8(1), 73-87. https://doi.org/10.1108/JTF-07-2019-0065

Saura, J.R. (2021). Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics. *Journal of Innovation & Knowledge*, 6(2), 92–102. https://doi.org/10.1016/j.jik.2020.08.001

Shaik, M. (2023). Impact of artificial intelligence on marketing. *East Asian Journal of Multidisciplinary Research*, 2(3), 993–1004. https://doi.org/10.55927/eajmr.v2i3.3112

Shahid, M.Z., & Li, G. (2019). Impact of artificial intelligence in marketing: a perspective of marketing professionals of Pakistan. Global *Journal of Management and Business Research*, 19(2), 27-33. https://journalofbusiness.org/index.php/GJMBR/article/view/2704

Telikani, A., Tahmassebi, A., Banzhaf, W., & Gandomi, A.H. (2021). Evolutionary Machine Learning: A Survey. *ACM Computing Surveys*, 54(8), 1–35. https://doi.org/10.1145/3467477

Theodoridis, P.K., & Gkikas, D.C. (2019). How artificial intelligence affects digital marketing. In Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018. *Springer International Publishing*, 151. https://doi.org/10.1007/978-3-030-12453-3 _

Thiraviyam, T. (2018). Artificial intelligence marketing. *International Journal of Recent Research Aspects*, 4, 449-452. https://doi.org/10.1016/j.jjimei.2020.100002

Volkmar, G., Fischer, P.M., & Reinecke, S. (2022). Artificial Intelligence and Machine Learning: Exploring drivers, barriers, and future developments in marketing management. *Journal of Business Research*, 149, 599–614. https://doi.org/10.1016/j.jbusres.2022.04.007

Yang, H., Nepusz, T., & Paccanaro, A. (2012). Improving GO Semantic Similarity Measures by Exploring the Ontology Beneath the Terms and Modeling Uncertainty. *Bioinformatics*, 28(10), 1383–1389

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means *Clustering Algorithm*. J, 2(2), 226–35. https://doi.org/10.3390/j2020016

Zhao, S., Li, Z., Huang, X., Rupp, A., Göser, J., Vovk, I.A., Kruchinin, S.Y., Watanabe, K., Taniguchi, T., Bilgin, I., Baimuratov, A.S., & Högele, A. (2022). Excitons in mesoscopically reconstructed moiré heterostructures. *Bioinformatics.* https://doi.org/10.48550/arXiv.2202.11139

Zou, J., Kanoulas, E., & Liu, Y. (2020). An Empirical Study on Clarifying Question-Based Systems. *Bioinformatics*, 28(10), 1383–1389. https://doi.org/10.1145/3340531.3412094

Article history: Received: 24.01.2024 Revised: 04.04.2024 Accepted: 25.04.2024 Available online: 23.05.2024