engineering proceedings

MDPI

Proceeding Paper

Exploring Regional Determinants of Tourism Success in the Eurozone: An Unsupervised Machine Learning Approach †

Charalampos Agiropoulos [1,*], James Ming Chen [2], George Galanos [1] and Thomas Poufinas [3]

[1] Department of International and European Studies, University of Piraeus, 185 34 Pireas, Greece; georgegalanos3@gmail.com
[2] College of Law, Michigan State University, East Lansing, MI 48824, USA; chenjame@law.msu.edu
[3] Department of Economics, Democritus University of Thrace, 691 00 Komotini, Greece; tpoufinas@gmail.com
* Correspondence: hagyropoylos@gmail.com
† Presented at the 10th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 15–17 July 2024.

Abstract: This paper presents an initial analysis of the factors influencing tourism success at the NUTS 2 regional level across the Eurozone from 2010 to 2019. Utilizing an extensive dataset that includes economic, demographic, and tourism-specific indicators, we employ unsupervised machine learning techniques, primarily K-means clustering and Principal Component Analysis (PCA), to unearth underlying patterns and relationships. Our study reveals distinct clusters of regions characterized by varying degrees of economic prosperity, infrastructure development, and tourism activity. Through K-means clustering, we identified optimal groupings of regions that share similar characteristics in terms of GDP per capita, unemployment rates, tourist arrivals, and overnight stays, among other metrics. Subsequent PCA provided deeper insights into the most influential factors driving these clusters, offering a reduced-dimensional perspective that highlights the primary axes of variation. The findings underscore significant disparities in tourism success across the Eurozone, with economic robustness and strategic infrastructural investments emerging as key drivers. Regions with higher GDP per capita and lower unemployment rates tend to exhibit higher tourism metrics, suggesting that economic health is a substantial contributor to regional tourism appeal and capacity. This paper contributes to the literature by demonstrating how machine learning can be applied to regional tourism data to better understand and strategize for tourism development. The insights garnered from this study are poised to assist policy-makers and tourism planners in crafting targeted interventions aimed at enhancing tourism competitiveness in underperforming regions.

Keywords: tourism determinants; unsupervised machine learning; K-means; regional analysis

JEL Classification: L83; O52; R11; C38; Z32

Citation: Agiropoulos, C.; Chen, J.M.; Galanos, G.; Poufinas, T. Exploring Regional Determinants of Tourism Success in the Eurozone: An Unsupervised Machine Learning Approach. *Eng. Proc.* **2024**, *68*, 53. https://doi.org/10.3390/engproc2024068053

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 19 July 2024

1. Introduction

Tourism stands as a multifaceted economic contributor, pivotal to the socioeconomic fabric of the Eurozone [1]. The synergy of global connectivity and localized cultural richness makes the analysis of regional tourism dynamics not only intellectually enriching but also essential for informed policy-making [2]. This is particularly pertinent to the NUTS 2 regions, where the heterogeneity in economic, cultural, and environmental attributes requires a nuanced approach to understanding tourism determinants [3].

Within the Eurozone, tourism is a significant economic driver, with varied impacts across its diverse regions [4]. The decade from 2010 to 2019 presents a unique temporal canvas marked by both economic challenges, such as the sovereign debt crisis, and a significant growth in global travel, presenting an opportunity to explore the underlying factors that influence tourism at a regional level [5]. This period's rich data, encapsulated

in our panel data time series, provide a foundation for a detailed empirical analysis of tourism determinants.

To provide a nuanced understanding of the interplay between economic factors and tourism development, this study employs K-means clustering analysis. This powerful statistical technique allows for the identification of distinct groupings (clusters) of Eurozone regions based on their similarities in economic indicators, demographic factors, and tourism outputs [6]. By delineating such clusters, this analysis reveals the specific combinations of characteristics that are associated with varying levels of tourism success.

The transformation of human flow into tourism revenue is an intricate process, influenced by a region's capacity to accommodate, entertain, and engage tourists [7]. The Eurozone's NUTS 2 regions, with their distinct characteristics, offer a rich setting to explore these dynamics [8]. This study's methodological breadth enables a nuanced exploration of these regions' capacity to convert potential and actual human traffic into economic gain, thereby contributing to the burgeoning literature on sustainable tourism development [9].

The objectives of this research are twofold: to identify and analyze the determinants that statistically influence tourism flows within the Eurozone, and to offer practical insights for regional policy-makers to leverage tourism as an economic development tool [10]. The anticipated outcome is a data-driven framework that aligns with the strategic imperatives of economic resilience and cultural sustainability [11].

The remainder of this paper proceeds as follows. First, the methodology section details the data sources, variables, and the K-means clustering approach employed in the analysis. Next, the results section presents the distinct cluster profiles that emerged from the analysis, highlighting their key economic, demographic, and tourism characteristics. The discussion section examines the theoretical and policy implications of the findings, considering potential strategies for tourism development in less prosperous regions and exploring the potential for broader economic growth catalyzed by the tourism sector. Finally, the paper concludes by outlining limitations of the study and suggesting avenues for future research.

2. Literature Review and Background

Tourism is an integral component of the economic structure within the Eurozone, having substantial implications for regional development and employment [6]. The significance of tourism's economic impact is particularly pronounced in the NUTS 2 regions, which exhibit a wide variety of cultural and natural attractions [3]. This literature review synthesizes key findings on the determinants of tourism in these regions, outlining economic, socio-cultural, and infrastructural factors, as well as methodological approaches employed in previous studies.

2.1. Economic and Socio-Cultural Determinants

The economic factors influencing tourism include affordability, exchange rates, and economic stability [12] emphasize the role of economic policy in shaping the tourism landscape, highlighting the need for fiscal measures that support the industry. Moreover, ref. [13] suggest that economic cycles significantly impact tourism, necessitating a detailed understanding of macroeconomic influences.

Cultural and heritage elements are equally vital in attracting tourists. Ref. [14] points out that regions rich in cultural heritage sites often see higher tourist numbers, supporting the idea that cultural capital is a significant draw for visitors. In addition, culinary experiences and language accessibility play a role in enhancing tourist satisfaction and prolonging stays [15].

2.2. Infrastructure and Regional Capacity

Infrastructure's role in tourism, particularly transportation and accommodation, has been well documented. Ref. [9] argues that infrastructure not only facilitates access to regions but also affects the overall tourist experience, impacting decisions to return. Effective

transportation systems and a range of accommodation options are crucial for converting tourist interest into actual visits [16].

The concept of regional capacity, which encompasses the ability of a region to sustain tourism without degrading its natural and cultural assets, is increasingly recognized as important for long-term tourism development [17]. Ref. [15] delves into the conversion of human traffic into revenue, underscoring the need for regions to manage capacity to maintain competitiveness and sustainability.

2.3. Methodological Approaches in Tourism Research

The diverse nature of tourism, spanning economic, social, cultural, and environmental dimensions, necessitates a multi-pronged methodological toolkit. Researchers utilize a spectrum of approaches, broadly categorized into quantitative, qualitative, and mixed-methods designs.

2.3.1. Quantitative Approaches

- Statistical Analysis: Statistical methods are widely employed to analyze large-scale tourism datasets, enabling the identification of trends, correlations, and the modeling of demand patterns [18]. Techniques such as regression analysis are used to examine the relationships between tourism outcomes (e.g., tourist arrivals, expenditure) and various determinants (e.g., economic indicators, marketing efforts, infrastructure).
- Economic Modeling: Tools like input–output analysis and computable general equilibrium (CGE) models help assess both the direct and indirect economic impacts of tourism on regional and national economies [19].
- Surveys: Structured surveys offer a means to collect quantifiable data on tourists' demographics, travel behavior, motivations, preferences, and satisfaction levels [13].

2.3.2. Qualitative Approaches

- Interviews: In-depth interviews (semi-structured or unstructured) provide rich insights into tourists' experiences, the perspectives of tourism stakeholders, and the lived realities of communities impacted by tourism [20].
- Focus Groups: Focus groups allow for the exploration of collective opinions, shared experiences, and the dynamics of social interaction within tourism contexts [21].
- Ethnography and Participant Observation: Researchers immerse themselves in tourism settings to gain a deep understanding of cultural practices, social interactions, and the power dynamics surrounding tourism activities [22].
- Content Analysis: Discourse analysis and other content analysis techniques are used to examine textual and visual representations in tourism marketing, policy documents, or online reviews, to expose underlying narratives and power structures [23].

2.3.3. Mixed-Methods Approaches

The integration of quantitative and qualitative methods offers a more holistic understanding of complex tourism phenomena [24]. For instance, a mixed-methods study might combine survey data on tourist satisfaction with interviews to delve into the qualitative aspects of the tourist experience.

2.4. Machine Learning Technigues the Tourism Sector

Machine learning has become an increasingly valuable tool within the tourism sector, offering data-driven insights to optimize operations and enhance the travel experience. Much of the existing research focuses on supervised learning methods. For instance, studies utilize regression and classification models to forecast tourism demand based on historical trends and economic indicators [18,25]. Supervised techniques also prove effective in customer segmentation, allowing tourism businesses to tailor marketing strategies and personalize offerings.

Despite the success of supervised learning, the potential of unsupervised machine learning techniques in the tourism context remains relatively under-explored. Unsupervised methods excel at identifying hidden patterns and groupings within complex datasets without the need for pre-defined labels. Clustering algorithms, for example, have been employed to segment tourists based on behavioral data, revealing niche target audiences and uncovering trends that may go unnoticed by traditional segmentation techniques [26]. Similarly, topic modeling, a text analysis technique, has been used to analyze online reviews and identify key themes shaping tourists' perceptions of destinations [27].

The limited body of work focusing on unsupervised machine learning in tourism highlights a significant research gap. This gap warrants further investigation, considering unsupervised methods' ability to address unique tourism challenges. Anomaly detection could aid in identifying fraudulent activity or unusual patterns within tourism data, ensuring security and mitigating risks [28]. Moreover, as tourism businesses increasingly seek to personalize experiences, unsupervised techniques are well-positioned to reveal nuanced patterns in tourist preferences that may remain hidden in supervised learning approaches.

3. Data and Methodology

3.1. Data Description

This study utilizes an extensive dataset that includes multiple indicators of tourism, economic performance, and demographic characteristics at the NUTS 2 regional level across the Eurozone from 2010 to 2019. The dataset comprises variables such as population, average population age, hospital bed availability, heating and cooling usage, unemployment rates, GDP per capita, tourist arrivals, tourism establishments, overnight stays, and a regional competitiveness index. These variables were chosen to provide a comprehensive view of the factors that might influence the tourism success of a region.

3.2. Methodological Framework

3.2.1. Unsupervised Machine Learning

To uncover the underlying patterns and structures within the data, we employed unsupervised machine learning techniques, specifically K-means clustering and Principal Component Analysis (PCA). These methods were chosen for their ability to classify data without prior labels and to reduce dimensionality, respectively.

3.2.2. K-Means Clustering

K-means clustering was applied to identify inherent groupings within the regions based on their characteristics. The optimal number of clusters was determined using the elbow method, which involved calculating the sum of squared distances from each point to its assigned center as the number of clusters varied. This method indicated a clear bend at four clusters, suggesting this as the optimal number of groupings for our analysis.

The clustering process was executed using the K-means class from scikit-learn in Python, specifying four clusters, and initializing the K-means algorithm with a random state for reproducibility. Each region was then assigned to one of the four clusters based on its characteristics, with subsequent analysis providing insights into the economic and tourism dynamics within each cluster.

3.2.3. Principal Component Analysis (PCA)

PCA was conducted to reduce the dimensionality of the data and to uncover the principal components that capture the most significant variance within the dataset. This technique transformed the high-dimensional data into a new coordinate system with dimensions (principal components) ordered by the variance they capture from the original data.

The PCA implementation from scikit-learn was utilized, with an initial analysis involving all components to observe the cumulative variance explained by successive components. The first few components that accounted for approximately 80% to 90% of the variance were considered sufficient for capturing the essential characteristics of the data. This

reduced-dimensional data facilitated a deeper understanding of the factors driving regional differences in tourism success.

3.3. Data Processing

The data were preprocessed to ensure optimal outcomes from the applied machine learning techniques. This involved handling missing values, normalizing data to ensure that variables with larger scales do not unduly influence the results, and encoding categorical variables where necessary. The normalization was particularly important for the application of K-means clustering, as this algorithm uses Euclidean distance between points which can be disproportionately affected by the scale of the data.

3.4. Analytical Procedures

This analysis employed a multi-stage approach to investigate tourism patterns across European regions. The following steps were involved:

I. Exploratory Data Analysis (EDA): Initially, the data were carefully examined to understand their underlying distributions, identify potential outliers, and visualize relationships between variables. This step likely involved techniques such as summary statistics (means, medians, standard deviations), histograms, scatterplots, and correlation matrices.
II. Clustering Implementation: The K-means clustering algorithm was applied to the preprocessed data. K-means is an unsupervised machine learning technique that groups data points based on their similarity. In this case, it would have been used to identify and classify European regions with similar tourism characteristics.
III. PCA Implementation: Principal Component Analysis (PCA) was utilized to further analyze the data structure and aid in visualization. PCA is a dimensionality reduction technique that transforms the original variables into a smaller number of 'principal components'. These principal components capture the majority of the variation within the data and allow for easier visualization of high-dimensional datasets. PCA likely helped to visualize the clusters identified by K-means and further understand the key factors driving the differences between European regions in terms of tourism.

4. Analysis

4.1. Exploratory Data Analysis

Exploratory Data Analysis was conducted as the initial phase of the study to gain a deeper understanding of the data characteristics and to uncover the underlying patterns within the variables. This stage involved several key activities:

- Distribution Analysis: We examined the distributions of key variables such as GDP per capita, unemployment rates, tourist arrivals, and overnight stays. This helped in identifying outliers, understanding the spread of data, and preparing for further cleaning and normalization.
- Correlation Analysis: Correlation matrices were generated to explore the relationships between different economic and tourism-related variables. This analysis was crucial to identify variables that strongly influence tourism success, such as the link between GDP per capita and overnight stays.
- Visual Exploration: Various visualizations including histograms, box plots, and scatter plots were used to visualize data distributions and relationships. For instance, scatter plots of GDP per capita versus overnight stays highlighted regions with potential underutilized tourism capacities despite economic prosperity.

Table 1 summarizes the descriptive statistics for the dataset, providing insights into the central tendencies, variability, and distributional shapes of the variables. Population exhibits a wide range (27,734–12.25 million) with a median of 1.45 million, suggesting a mix of smaller and larger regions. Age reveals demographic diversity with an average age range

of 33 to 51.7 years. Hospital beds demonstrate significant variation (49–70,948), highlighting disparities in healthcare infrastructure. Heating and cooling energy consumption also exhibit considerable variability, likely reflecting diverse climates and energy practices across the regions.

Table 1. Descriptive statistics of the tourism dataset.

	Population	Age	Hospital	Heating	Cooling	Unemployment	gdp_pc	Arrivals	Establishments	Nights	Competitiveness
count	1760	1760	1760	1760	1760	1760	1760	1760	1760	1760	1760
mean	1,922,569	43.19	10,329	2441	110	9686	30,124	2,493,439	129,791	6,645,061	−0.010
std	1,788,121	3.14	9661	960	151	6169	12,070	2,626,598	140,020	7,147,940	0.626
min	27,734	33.00	49	42	0	1000	8500	33,004	760	0	−1.610
25%	728,088	41.20	3284	1808	11	5400	22,000	803,741	47,739	2,111,003	−0.438
50%	1,453,361	43.50	8417	2562	36	7900	28,849	1,706,007	79,342	4,406,289	0.100
75%	2,338,724	45.30	14,118	2997	158	12,000	36,025	2,987,430	149,957	7,967,022	0.448
max	12,252,917	51.70	70,948	6508	812	37,000	102,200	21,828,739	794,251	40,670,263	1.360

Economic indicators show notable disparities. Unemployment rates range from 1% to 37% (mean of 9.69%), while GDP per capita varies significantly (8500–102,200). Tourism activity indicators (tourist arrivals and nights spent) also show wide ranges, suggesting uneven levels of tourism development across regions. The number of tourism-related establishments (760–794,251) further confirms this variation. Finally, the tourism competitiveness index (−1.61 to 1.36) indicates varying levels of regional competitiveness in attracting visitors.

The distribution plots in Figure 1 reveal several key characteristics of the variables under consideration. The hospital bed distribution exhibits a rightward skew, indicating a concentration of regions with lower bed counts and a smaller number of regions possessing significantly higher bed capacities. A similar right-skewness is observed for heating, with most regions demonstrating lower energy consumption and a few outliers exhibiting considerably higher consumption levels. The distribution for cooling demonstrates a pronounced rightward skew, with a peak at very low values. This suggests that cooling systems are either less prevalent or less intensively utilized across the majority of regions in comparison to heating systems. Finally, competitiveness appears to follow a somewhat bimodal distribution, implying a tendency for regions to cluster around distinct levels of competitiveness.

These distributional characteristics highlight the presence of variability and skewness within the dataset. This information is crucial for data preprocessing decisions prior to the application of machine learning techniques such as K-means clustering and PCA. Transformations may be necessary to normalize skewed distributions, ensuring optimal model performance and the generation of more insightful clustering results.

The analysis of the correlation matrix heatmap as shown in Figure 2 unveils several key relationships among the examined variables. A significant negative correlation (−0.54) exists between GDP per capita (gdp_pc) and unemployment, implying that regions exhibiting higher GDP per capita generally experience lower unemployment rates. As anticipated, tourist arrivals and nights spent demonstrate a strong positive correlation (0.98), confirming the direct relationship between these tourism metrics.

Population size and the number of tourism-related establishments exhibit a moderate positive correlation (0.64), suggesting an association between larger populations and increased tourism infrastructure. A similar moderate positive correlation (0.57) is observed between a region's competitiveness index and its GDP per capita. This indicates that greater economic prosperity tends to coincide with increased competitiveness within the tourism sector.

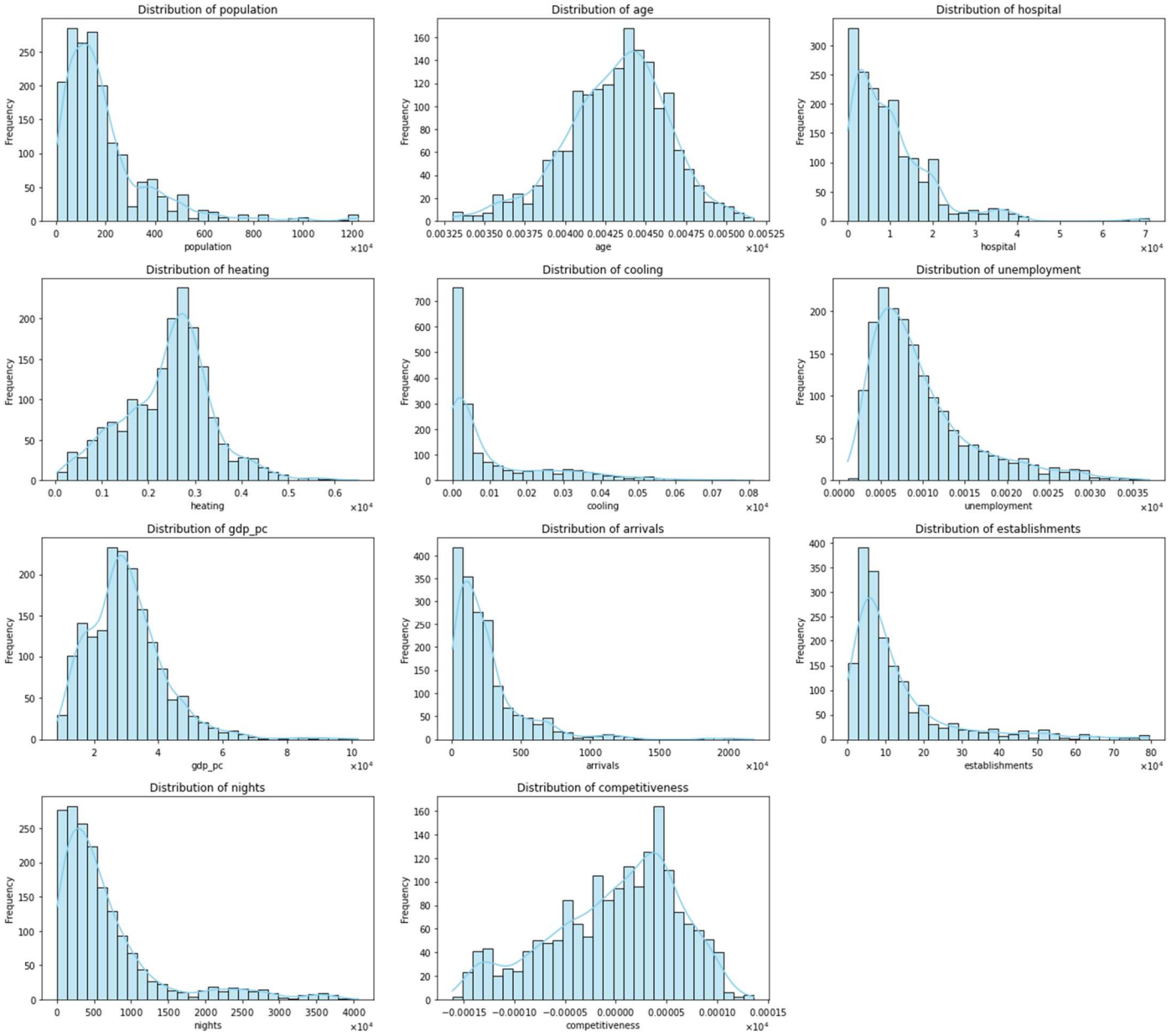


Figure 1. Histograms and distribution plots for all key variables.

Additionally, a positive correlation (0.77) links tourist arrivals and the number of establishments, demonstrating that regions possessing a greater number of tourism-related establishments typically attract more visitors.

Hospital beds show correlations with several variables, though these are generally weaker than those observed for indicators like GDP or population. This may suggest a connection between healthcare capacity and long-term tourism, potentially influencing a region's overall attractiveness. Heating and cooling energy consumption display some correlation with other metrics, potentially reflecting regional climate and infrastructure differences related to energy use.

Finally, the competitiveness index exhibits notable correlations with key economic and tourism variables, including GDP per capita, tourist arrivals, and nights spent. This underscores the index's value in assessing drivers of tourism success within a region.

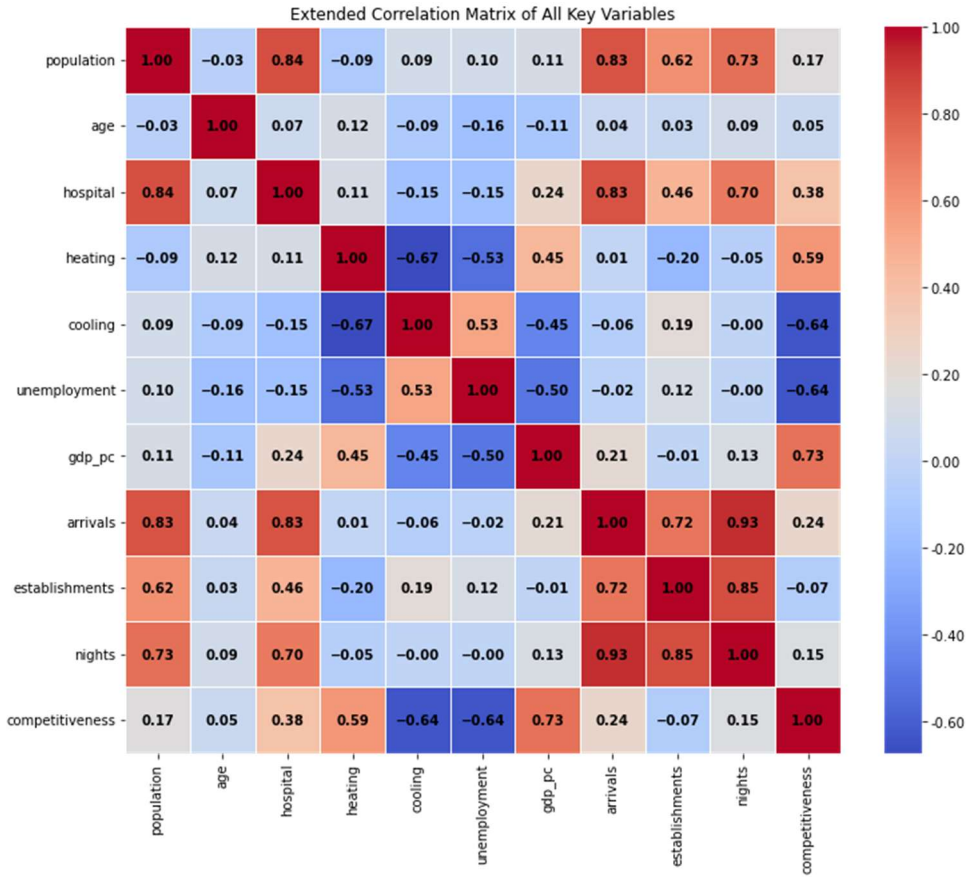


Figure 2. Correlation heatmap of all key variables.

4.2. K-Means Clustering

Figure 3 illustrates the variation in inertia (within cluster sum of squares) as a function of cluster count (k″), employing the elbow method for optimal cluster determination. The plot suggests that the curve exhibits a noticeable inflection point around k = 4. This implies that the optimal number of clusters for the dataset may lie in the vicinity of four. Beyond this point, the addition of clusters does not yield a substantial reduction in inertia, indicating decreasing marginal gains in cluster cohesion.

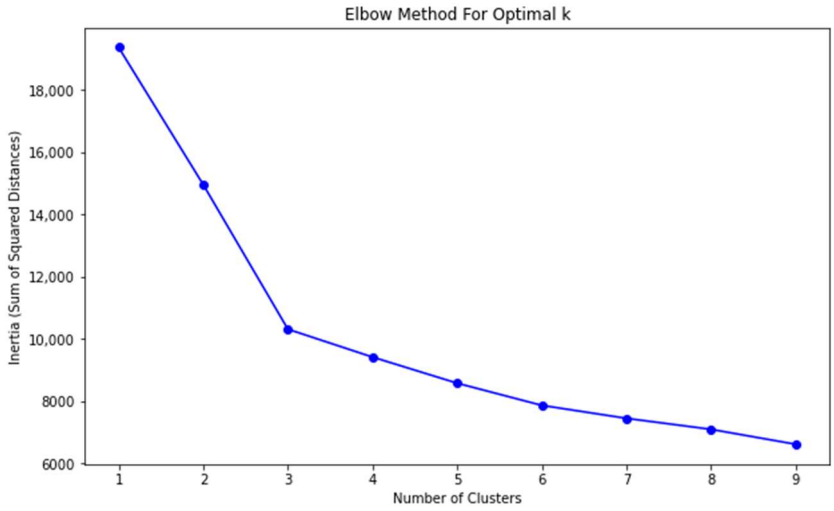


Figure 3. Elbow method for optimal cluster determination.

To partition the NUTS 2 regions according to their tourism and socioeconomic profiles, K-means clustering was implemented with a 4-cluster solution. Prior to analysis, data normalization was performed. This step ensures that variables are compared on an equitable scale and that distance calculations within the clustering algorithm are not biased by differing units of measurement.

4.3. Cluster Characteristics

The K-means cluster analysis revealed four distinct profiles of Eurozone regions (Table 2), each characterized by unique combinations of economic and tourism indicators:

- Cluster 0 (High Economic Prosperity, Strong Tourism): This cluster is characterized by high GDP per capita and low unemployment rates. Regions within this cluster demonstrate significant tourism activity and possess high competitiveness scores within the tourism sector.
- Cluster 1 (Economic Challenges, Limited Tourism): Regions in this cluster exhibit the lowest GDP per capita and the highest unemployment rates. These factors align with lower tourist arrivals, nights spent, and weaker tourism competitiveness indices.
- Cluster 2 (High Economic Prosperity, Moderate Tourism): This cluster boasts the highest GDP per capita but displays moderate levels of tourism activity. Despite having fewer tourism-related establishments and nights spent compared to Cluster 0, these regions maintain strong competitiveness scores.
- Cluster 3 (Tourism Focus, Moderate Competitiveness): This cluster is composed of highly populous regions experiencing significant tourist arrivals and nights spent. These areas represent major tourism destinations; however, their competitiveness scores are relatively modest when compared to their economic scale.

Table 2. Summarized table reflecting the average characteristics of each cluster.

Cluster	Population	Age	Hospital Beds	Heating	Cooling	Unemployment (%)	GDP per Capital	Tourist Arrivals	Establishments	Nights Spent	Competitiveness
0	1,347,321	42.69	4715	1369	273	15.86	18,467	1,210,073	110,257	3,604,852	−0.82
1	1,331,712	43.24	8334	2996	29	6.76	34,961	1,743,397	76,715	4,305,789	0.32
2	3,776,541	44.4	21,861	2357	103	8.67	32,113	5,533,635	271,811	15,814,081	0.21
3	8,088,220	41.21	37,690	2182	164	12.97	34,469	11,742,114	538,763	29,445,775	0.13

4.4. Visual Analysis

Scatter plots in Figure 4 and bar graphs in Figure 5 were used to visualize the clusters in terms of key metrics such as GDP per capita vs. tourist arrivals and population vs. competitiveness. These visual representations helped clarify the distinctions between clusters, providing a visual confirmation of the cluster analysis.

4.4.1. GDP per Capita vs. Tourist Arrivals

The analysis of GDP per capita versus tourist arrivals reveals interesting patterns across the clusters. Within Cluster 3 (major tourism hubs), regions exhibit consistently high tourist arrivals despite moderate variability in GDP per capita. This suggests that robust tourist inflow may not be strictly dependent upon the highest levels of economic prosperity within this cluster. Clusters 0 and 2, characterized by economic strength, demonstrate a generally positive correlation between GDP per capita and tourist arrivals. However, the range of variation in tourist arrivals within these clusters is less pronounced compared to Cluster 3. Finally, Cluster 1, marked by economic challenges, exhibits consistently lower levels of both GDP per capita and tourist arrivals, underscoring the association between economic standing and tourism performance.

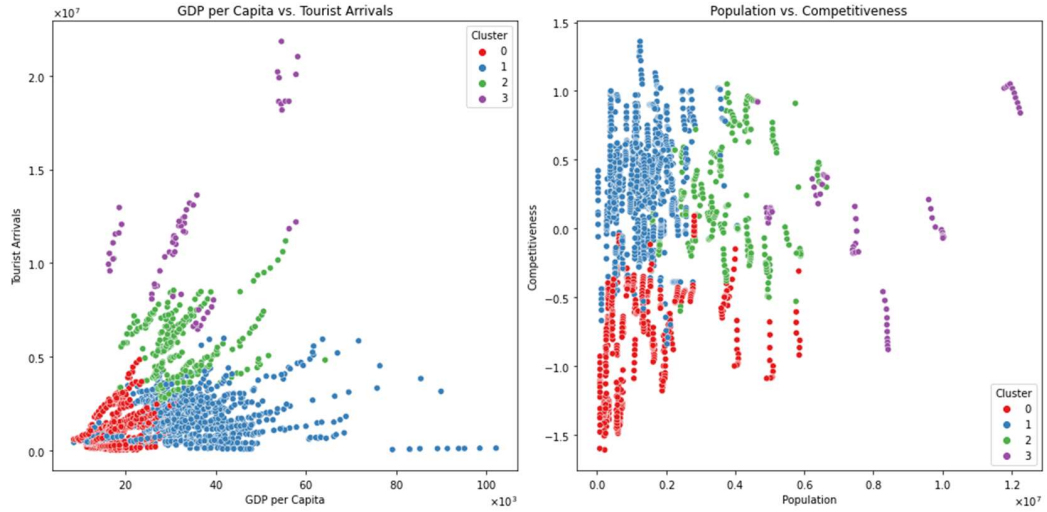
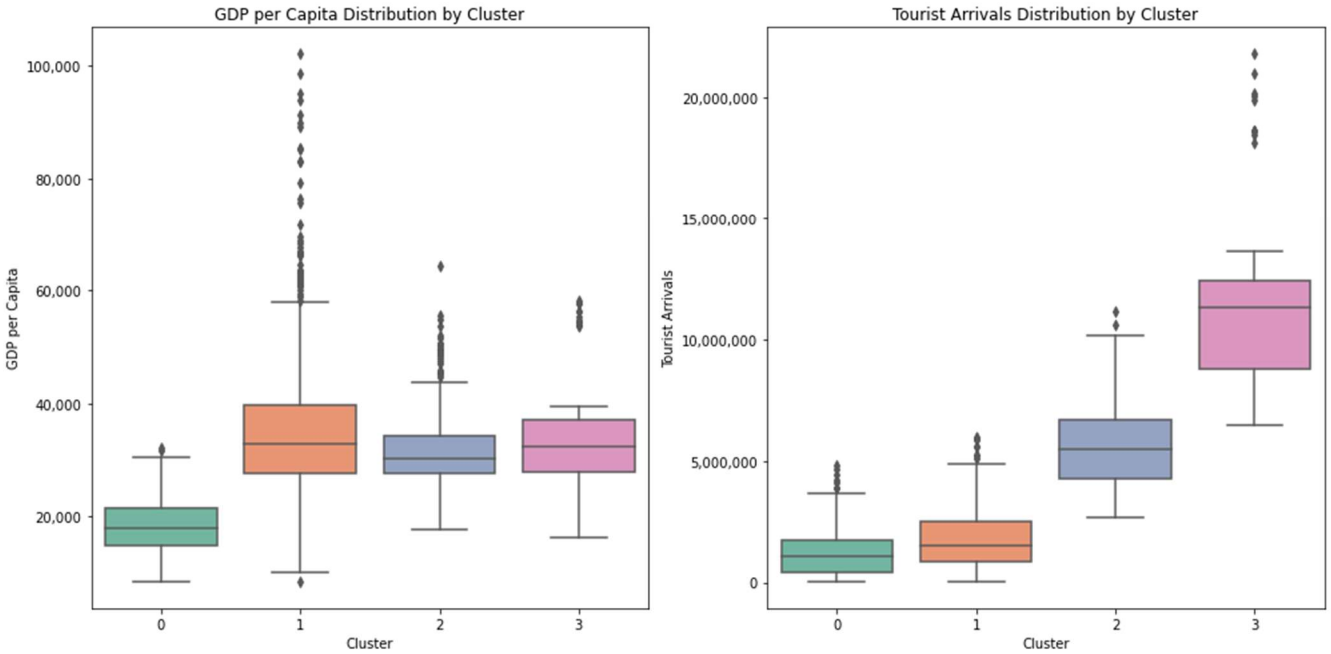


Figure 4. Indicative scatter plots for GDP per capita vs. tourist arrivals and population vs. competitiveness.



(**a**) (**b**)

Figure 5. Indicative box plots for GDP per capita (**a**) and tourist arrivals (**b**) by cluster.

4.4.2. Population vs. Competitiveness

The relationship between population and competitiveness presents a more nuanced picture. Regions in Cluster 3, despite their large populations, demonstrate moderate competitiveness scores. This observation suggests that while these regions successfully attract significant tourist numbers, they may possess further potential to optimize their competitiveness relative to their demographic scale. In contrast, Clusters 0 and 2 generally exhibit higher competitiveness scores. Interestingly, these clusters tend to have smaller populations compared to Cluster 3, potentially indicating a link between population size and the ability to maximize tourism competitiveness.

4.4.3. Distribution of GDP per Capita across Clusters

The distribution of GDP per capita aligns with the previously established economic profiles of the clusters. Clusters 0 and 2 consistently exhibit higher GDP per capita values. Interestingly, Cluster 2's wider distribution suggests some internal variability in levels of

economic prosperity. Cluster 1 demonstrates the lowest median GDP per capita, reinforcing the economic challenges faced by regions within this cluster. Cluster 3, while characterized by a large population and strong tourism metrics, exhibits a significant spread in its GDP per capita distribution. This indicates notable variation in economic wealth across these major tourism hubs.

4.4.4. Distribution of Tourist Arrivals across Clusters

Cluster 3, as expected, exhibits the most significant tourist arrivals, with both a high median and a wide interquartile range reflecting its major tourism hub status. Cluster 0 also demonstrates substantial tourist activity, albeit less pronounced than Cluster 3. This aligns with its profile of economic strength and moderate population size. Clusters 2 and 1 display lower and more tightly clustered distributions of tourist arrivals. This pattern corresponds to their smaller economic scales and lower competitiveness scores within the tourism sector (Figure 5b).

4.5. Principal Components Analysis (PCA)

To further investigate the complex relationships within our dataset, we employ Principal Component Analysis (PCA). This powerful dimensionality reduction technique allows us to identify the most influential variables driving the observed patterns, while also visualizing the inherent structure of the data. PCA achieves this by creating new, uncorrelated variables (principal components) as linear combinations of the original variables. These principal components capture the maximum possible variance within the dataset in decreasing order of importance.

Figure 6 presents the cumulative variance explained by the PCA components and indicates that the first four components capture approximately 80% of the total variance in the dataset. To capture an even greater proportion of the variance, up to six components would be necessary, explaining roughly 90% of the total variance. This analysis suggests that using the first four to six principal components for further analyses would likely be sufficient to retain the most important information from the original dataset, while achieving a substantial reduction in dimensionality. This dimensionality reduction can be particularly advantageous for visualization and further statistical modeling techniques.

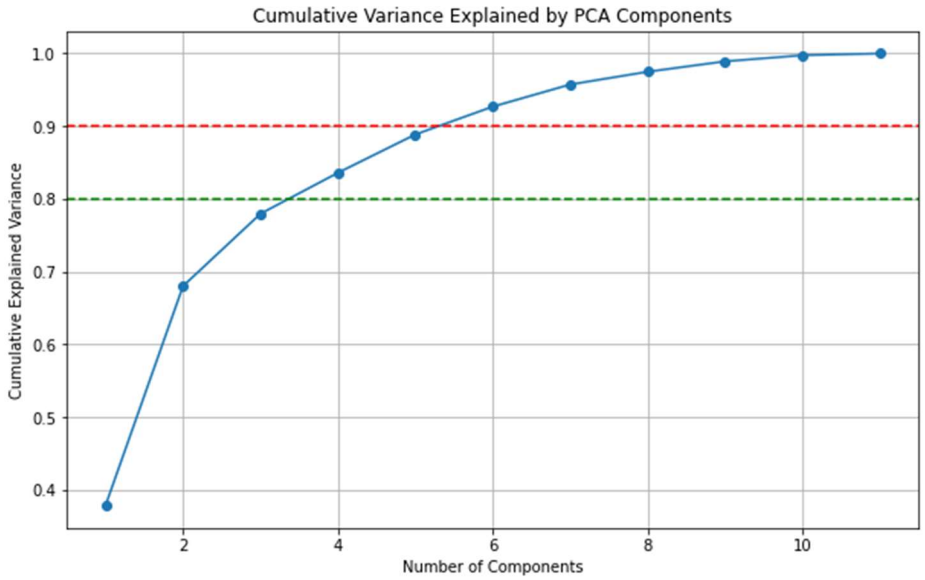


Figure 6. Cumulative variance explained using PCA components.

The visualization of the first two principal components reveals distinct groupings of regions, demonstrating that the PCA has successfully identified the primary dimensions of variability that discriminate between the previously defined clusters (Figure 7). The spatial

distribution of regions associated with each cluster suggests that the principal components effectively convey their distinguishing characteristics. Differences in cluster spread within the plot may suggest varying degrees of internal homogeneity. For example, a cluster significantly extended along the first principal component indicates that this component explains substantial variance relevant to the dominant features characterizing that cluster.

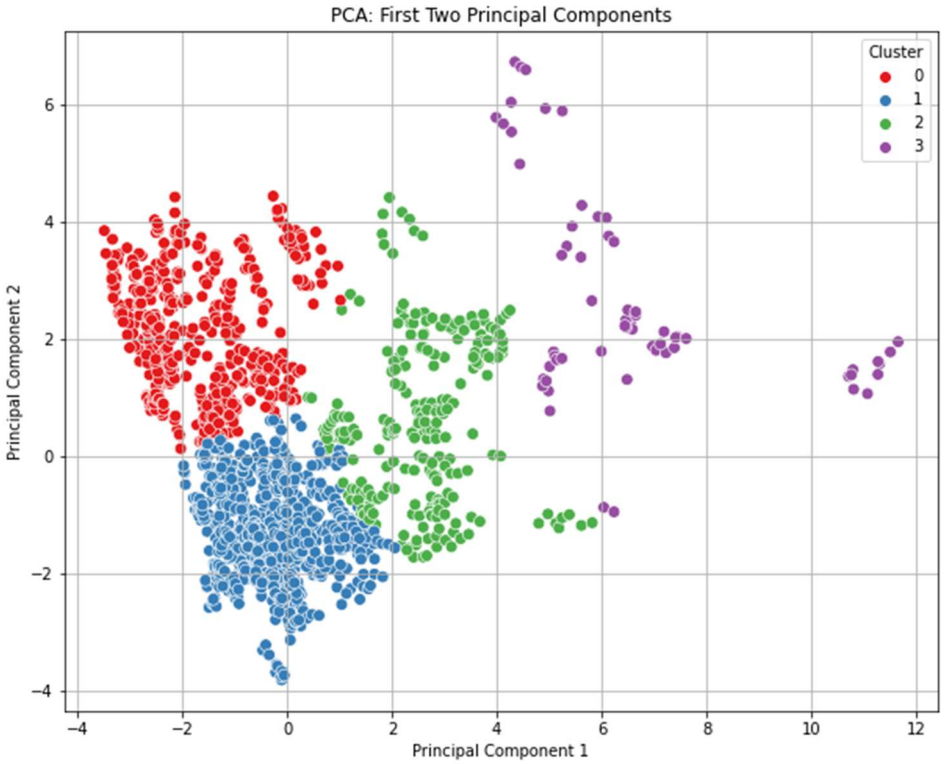


Figure 7. Plot of the first two principal components.

The interpretation of the principal components is crucial. Principal Component 1 (PC1) appears to capture variance associated with a region's overall economic well-being and tourism scale. This is evidenced by its ability to differentiate clusters based on metrics such as GDP and tourist activity. Principal Component 2 (PC2) may reflect factors related to infrastructural development or competitiveness, given its role in vertically delineating clusters within the plot.

5. Discussion

This study employed unsupervised machine learning techniques, specifically K-means clustering and Principal Component Analysis (PCA), to explore the underlying patterns and influential factors affecting tourism success across Eurozone regions at the NUTS 2 level, using comprehensive economic, demographic, and tourism-specific indicators from 2010 to 2019. The results elucidate significant disparities in tourism success, driven by variances in economic robustness, infrastructure development, and strategic investments [29].

5.1. Clustering Insights

The K-means clustering revealed four distinct clusters of regions, each characterized by unique attributes influencing their tourism competitiveness and economic landscape. Cluster 0 encompasses regions with high GDP per capita and relatively low unemployment rates, paired with substantial tourist arrivals and overnight stays, signaling robust economic and tourism sectors. Conversely, Cluster 1 includes regions struggling economically, evident from the lowest GDP per capita and highest unemployment rates among the clusters, correlating with lower tourist arrivals and competitiveness.

Clusters 2 and 3, while economically diverse, highlight interesting dichotomies in tourism dynamics. Cluster 2, despite its high GDP per capita, shows only moderate levels of tourist activity, suggesting underutilized potential or possibly a focus on niche or luxury tourism markets. Cluster 3 represents major tourism hubs characterized by high population densities and significant tourist volumes, yet with moderate competitiveness, potentially indicating inefficiencies or saturation effects that could be mitigating higher potential value generation from tourism activities.

5.2. PCA Findings

PCA further supported these distinctions by reducing dimensionality to identify the most influential factors driving regional variations. The analysis of the first three principal components accounted for a significant proportion of the variance and underscored the multidimensional nature of regional tourism success. Notably, the first component appeared to capture economic size and capacity, the second linked closely with infrastructure and service quality, and the third potentially reflected variations in policy effectiveness or market saturation.

These components help in understanding the nuanced interactions between economic indicators and tourism metrics, offering a macroscopic view of how regions can leverage or enhance specific aspects of their tourism and economic structures to foster growth and competitiveness.

5.3. Implications for Policy and Practice

The findings provide critical insights for policy-makers and tourism planners aiming to enhance regional competitiveness in the tourism sector. For struggling regions (Cluster 1), focused interventions on improving economic conditions and enhancing basic tourism infrastructure could be vital. In contrast, regions in Cluster 2 might benefit from more targeted marketing strategies or the development of specialized tourism products that leverage their economic strengths without necessarily aiming for high visitor volumes [30].

For major tourism hubs (Cluster 3), strategies could involve the diversification of tourism offerings to spread visitor numbers more evenly throughout the year or enhancing tourist experiences to convert high visitor numbers into greater economic benefits [31]. The holistic understanding gained from this study assists in crafting nuanced, data-driven policies that cater to the specific needs and potentials of different regions.

5.4. Future Research Directions

Further research could explore longitudinal changes beyond 2019 to assess the impact of significant global events, such as the COVID-19 pandemic, on regional tourism dynamics. Additionally, integrating qualitative data on tourist satisfaction or regional brand image could enrich the quantitative analyses, offering deeper insights into the qualitative factors that influence tourism success.

5.5. Limitations of the Study

While the study provides valuable insights, there are limitations to consider. The reliance on quantitative data may overlook qualitative factors such as regional brand image, tourist satisfaction, and local community support, which can significantly influence tourism success. Future studies incorporating mixed methods could provide a more comprehensive understanding of the dynamics at play.

Author Contributions: Conceptualization, C.A. and J.M.C.; methodology, C.A. and J.M.C.; software, C.A.; validation, G.G., T.P. and J.M.C.; formal analysis, C.A.; investigation, C.A.; resources, G.G. and T.P.; data curation, C.A.; writing—original draft preparation, C.A.; writing—review and editing, C.A.; visualization, C.A.; supervision, C.A.; project administration, C.A.; funding acquisition, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH—CREATE—INNNot NOVATE (project SlotHub, code:T2EDK02208).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This research did not involve human subjects.

Data Availability Statement: Data was drawn from public sources. The authors are happy to provide their own curated version of that data upon request.

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

References

1. Antonakakis, N.; Dragouni, M.; Filis, G. How strong is the linkage between tourism and economic growth in Europe? *Econ. Model.* **2015**, *44*, 142–155. [CrossRef]
2. Hall, C.M. A typology of governance and its implications for tourism policy analysis. In *Tourism Governance*; Routledge: London, UK, 2013; pp. 27–47.
3. Eurostat. Regions in Europe–2022 Interactive Edition. 2022. Available online: https://ec.europa.eu/eurostat/cache/digpub/regions/#total-population (accessed on 8 July 2024).
4. Anagnostou, A.; Ekonomou, G.; Kallioras, D. The nexus of tourism spending with economic performance: A panel data analysis for the Eurozone area. *Geogr. Pannonica* **2021**, *25*, 35–44. [CrossRef]
5. Shpak, N.; Muzychenko-Kozlovska, O.; Gvozd, M.; Sroka, W. Simulation of the influence of external factors on the level of use of the regional tourism potential: A practical aspect. *Adm. Sci.* **2021**, *11*, 85. [CrossRef]
6. Bulin, D. EU Travel and Tourism Industry—A Cluster Analysis of Impact and Competitiveness. *Glob. Econ. Obs.* **2014**, *2*, 150–162.
7. Miller, G. The development of indicators for sustainable tourism: Results of a Delphi survey of tourism researchers. *Tour. Manag.* **2001**, *22*, 351–362. [CrossRef]
8. Aumayr, C.M. European Region Types: A Cluster Analysis of European NUTS 3 Regions. Ph.D. Dissertation, Vienna University of Economics and Business, Vienna, Austria, 2006.
9. Streimikiene, D.; Svagzdiene, B.; Jasinskas, E.; Simanavicius, A. Sustainable tourism development and competitiveness: The systematic literature review. *Sustain. Dev.* **2021**, *29*, 259. [CrossRef]
10. Stevenson, N.; Airey, D.; Miller, G. Tourism Policy Making:: The Policymakers' Perspectives. *Ann. Tour. Res.* **2008**, *35*, 732–750. [CrossRef]
11. Espiner, S.; Orchiston, C.; Higham, J. Resilience and sustainability: A complementary relationship? Towards a practical conceptual model for the sustainability–resilience nexus in tourism. *J. Sustain. Tour.* **2017**, *25*, 1385–1400. [CrossRef]
12. Edgell, D.L.; Swanson, J.; Allen, M.D.; Smith, G. *Tourism Policy and Planning: Yesterday, Today, and Tomorrow*; Routledge: London, UK, 2008.
13. Schubert, S.F.; Brida, J.G.; Risso, W.A. The impacts of international tourism demand on economic growth of small economies dependent on tourism. *Tour. Manag.* **2011**, *32*, 377–385. [CrossRef]
14. Panzera, E.; de Graaff, T.; de Groot, H.L. European cultural heritage and tourism flows: The magnetic role of superstar World Heritage Sites. *Pap. Reg. Sci.* **2021**, *100*, 101–122. [CrossRef]
15. Miller, G.; Rathouse, K.; Scarles, C.; Holmes, K.; Tribe, J. Public understanding of sustainable tourism. *Ann. Tour. Res.* **2010**, *37*, 627–645. [CrossRef]
16. Buhalis, D.; Amaranggana, A. Smart tourism destinations enhancing tourism experience through personalisation of services. In *Information and Communication Technologies in Tourism 2015: Proceedings of the International Conference, Lugano, Switzerland, 3–6 February 2015*; Tussyadiah, I., Inversini, A., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 377–389.
17. Robinson, M.; Novelli, M. Niche tourism: An introduction. In *Niche Tourism*; Routledge: London, UK, 2005; pp. 1–11.
18. Song, H.; Li, G. Tourism demand modelling and forecasting: A review of recent research. *Tour. Manag.* **2008**, *29*, 203–220. [CrossRef]
19. Dwyer, L.; Forsyth, P.; Spurr, R. Evaluating tourism's economic effects: New and old approaches. *Tour. Manag.* **2004**, *25*, 307–317. [CrossRef]
20. Jennings, G. *Tourism Research*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2010.
21. Krueger, R.A.; Casey, M.A. *Focus Groups: A Practical Guide for Applied Research*, 5th ed.; Sage: London, UK, 2015.
22. Phillimore, J.; Goodson, L. (Eds.) *Qualitative Research in Tourism: Ontologies, Epistemologies, and Methodologies*; Routledge: London, UK, 2004.
23. Stepchenkova, S.; Zhan, F. Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography. *Tour Manag.* **2013**, *36*, 590–601. [CrossRef]
24. Weaver, D.; Lawton, L. *Tourism Management*, 4th ed.; Wiley: Hoboken, NJ, USA, 2014; pp. 1–446.

25. Claveria, O.; Monte, E.; Torra, S. Tourism demand forecasting with neural networks models: Different ways of treating information. *Int. J. Tour. Res.* **2016**, *17*, 492–500. [CrossRef]
26. Dolnicar, S. A review of data-driven market segmentation in tourism. *J. Travel Tour. Mark.* **2002**, *12*, 1–22. [CrossRef]
27. Guo, Y.; Barnes, S.J.; Jia, Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tour. Manag.* **2017**, *59*, 467–483. [CrossRef]
28. Nunkoo, R.; Ramkissoon, H.; Gursoy, D. Use of structural equation modeling in tourism research: Past, present, and future. *J. Travel Res.* **2013**, *52*, 759–771. [CrossRef]
29. Formica, S.; Uysal, M. Destination attractiveness based on supply and demand perceptions. *Ann. Tour. Res.* **2006**, *33*, 418–430.
30. Novelli, M.; Schmitz, B.; Spencer, T. Networks, clusters and innovation in tourism: A UK experience. *Tour. Manag.* **2006**, *27*, 1141–1152. [CrossRef]
31. Dwyer, L.; Kim, C. Destination competitiveness: Determinants and indicators. *Curr. Issues Tour.* **2003**, *6*, 369–414. [CrossRef]