# Open-Data Mining in Tourism: Implications for Marketing Strategy

**PEDRO C. FLORES**[1] PhD., **FAUSTINO D. REYES**[2] II D.Tech., **AARON PAUL M. PINEDA**[3], DHRM, DBA
**MARILOU A. MADERAZO**[4], DBA

[1] Computer Science Division Ras Al Khaimah Campus, Higher Colleges of Technology, United Arab Emirates
[2] Information Technology Academy Bahrain Training Institute, Kingdom of Bahrain
[3] Business Division Al Dhafra Region Campuses, Higher Colleges of Technology, United Arab Emirates
[4] Business Division Ras Al Khaimah Campus, Higher Colleges of Technology, United Arab Emirates
Email: [1]pflores@hct.ac.ae, [2]faustino.reyes@bti.moe.bh, [3]apineda@hct.ac.ae, [4]mmaderazo@hct.ac.ae

**Abstract**

As technology advances, researchers will access new and advanced statistical techniques. This study examined the behavior of inbound visitors in the United Arab Emirates using data mining tools such as Bayanat, Statistica, and open government data. The study findings show customers' origins, nationalities, and visits by roughly compared observations. The researchers concluded that the biggest motivator for travelers' future visits to the UAE is not the experiences they had during their most recent visit but rather the experiences they expect to have in the future, such as visiting tourist attractions and seeing UAE improvements. The data mining approach almost eliminates researcher subjectivity. It allows for the valuable discovery of specific visitor patterns in massive data sets, giving governments and destination marketing organizations more tools to create effective destination marketing plans. The main goal of this study is to discuss and show data mining and its use in travel and tourism. This study aims to examine today's competitive environment for travel and tourism. To grow their market and maintain control, these organizations are being pushed to avoid using data mining tools and techniques to produce, manage, and advertise tourist products and services. The purpose of this study is to discuss and show data mining and its use in travel and tourism.

***Keywords**:* Open Data Mining, Business Planning, UAE Tourism, Tourism Trends

## 1. Introduction

Data is an integral element of any business venture. Every competitive business organization uses reliable and relevant data in making decisions. Hence, data has become a real asset as it crucially helps in determining the actions and directions every company wants to pursue. The correct, timely, and creative utilization of these data enables organizations to gain competitive and strategic advantage. Today, the development and democratization of business intelligence software empowers users even without deep-rooted technical expertise to analyze as well as extract insights from their data. As a result, less IT support is required to produce reports, trends, visualizations, and insights that facilitate the data decision making process.

The emergence of business intelligence tools that processes data, known as data analytics, stemmed from the fact that businesses now connect to or rely on IT systems such as cloud services, social media, and other online-based platforms. These systems generate massive and robust data wherein a business company can leverage information to make more informed and powerful business decisions. Several data analytics tools have the capability to find patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, one can derive useful information from data sets to increase revenues, cut costs, improve customer relationships, and reduce risks among others. Such technique is known as data mining. Moreover, mined data enables companies to create new business opportunities, predict future trends, optimize current operational efforts, and produce actionable insights.

The investigators of this study evaluated and utilized popular open source data mining tools to process and analyze datasets from three sources such as the UAE open data portal (www.bayanat.ae) site, Statistica portal, and UAE official tourism websites. The UAE open data portal is the official data portal of the UAE Government and is open for public access and use. The portal contains the datasets on economy, education, society, technology, transportation, environment, government, health and infrastructure. Statistica is a database German company which provides data statistics and services. Its website publishes up-to-date data of companies and governments worldwide. The investigators downloaded the tourism datasets through their account subscription to Statistica's database. The third dataset source includes Dubai tourism official websites. These websites provide periodic statistical report on the tourism of the country. Using the selected tools, the investigator conducted data mining processes on UAE open datasets. Based from the results of this study, there are relevant and useful information generated that can be utilized by strategists for business planning. Interestingly, the study shows that non-commercial data mining tools can be utilized by business companies and consultants to generate relevant information which may help and guide them in making strategic decisions.

**Statement of the Problem**

The central aim of this study was to explore and generate business insights through data mining open data projects related to UAE tourism.

Specifically, it sought to answer the following questions:

1) What data mining techniques generate relevant information to produce tourism business insights?

2) How does the COVID 19 pre-pandemic and post-pandemic tourism trends in the UAE compare?

**Conceptual Framework**

The main theme of this study was that data mining is an effective IT strategy to provide companies with meaningful business insights. Guided by this theme, this study utilized the three-layer conceptual framework proposed by Yao (2003) as illustrated in Fig. 1. The framework consists of the philosophy layer, the technique layer, and the application layer. The layered framework represents the understanding, discovery, and utilization of knowledge.

The philosophy layer investigates the basic issues of knowledge. The fundamental question on what knowledge and resources do we have is a critical input in this domain. There are many related issues to this question, such as the representation of knowledge, the expression and communication of knowledge in languages, the relationship between knowledge in mind and in the external real world, and the classification and organization of knowledge (Sowa, 1984). The philosophical study of data mining serves as an anchor to technology and application before these generate the knowledge and the understanding of our world. The philosophy layer study is primarily driven by curiosity, and responds to a certain hypothesis (Lin et al., 2005).

The technique layer is the study of knowledge discovery in machine. This domain attempts to answer how to discover knowledge. There are several issues related to this question, such as the implementation of human knowledge using tools which may require programming, storage and retrieval, techniques and algorithms for intelligent systems. Several studies across disciplines have

concentrated on the technique layer. Logical analysis and mathematical modeling are considered to be the foundations of technique layer of data mining (Lin, et al., 2005).

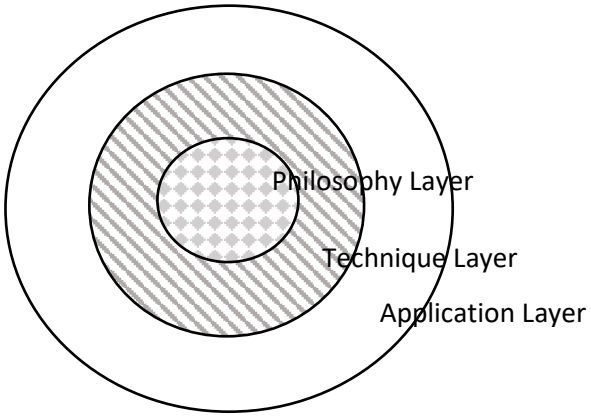


**Fig. 1 – Data Mining Three-Layer Framework**

The application layer focuses on the communication of results of the mined data to the target information consumers. The central goal is for the consumers to properly utilize the discovered knowledge. The tasks in this domain are crucial as the communicator, the last-mile team, should present meaningful analysis and visualization of the discovered knowledge. Meaningful analysis consists of repeated explorations as users develop insights about significant relationships, domain-specific contextual influences, and causal patterns. Visualization refers not only to a set of graphical images but also to the iterative process of visual thinking and interaction with the images. An interactive visualization environment, in which the user may choose to display the data in many different ways, encourages data exploration. One goal of data exploration is the recognition of pattern and the abstraction of structure and meaning from data (Minelli, et al., 2013).

**Literature Review**

According to Yuan (2022), big data is already gaining attention as it integrates with the tourism industry, offering up new avenues for tourism decision-making innovation. Tourists can search for travel information, book travel services, and share their travel experiences using mobile Internet technologies, resulting in "tourism big data." In particular, Karathiya (2012) stressed that data mining is the process of examining often large data sets to survey and discover previously unknown proto-types, styles, and relationships in order to generate knowledge for better decision making. Therefore, it has been determined that, in today's competitive market for travel and tourism, firms will be urged to use data mining tools and techniques to develop, manage, and advertise tourist products and services.

Furthermore, Hartmann (2021) mentioned that due to technology improvements, big data, and the availability of open source information, the business intelligence (BI) market has risen at an incredible rate in the last decade. Despite this expansion, the commercial sector's use of open government data (OGD) as a source of information is still limited due to a lack of understanding Contrary to this, in the study of Yan (2018), found out the following: 1) The use of open government data in research has steadily increased from 2009 to 2016; 2) Researchers use OGD from 96 different open government data portals, with http URL and this http URL being the most common sources; and 3. Evidence has provided that OGD from developing countries, particularly

India and Kenya, is frequently used to fuel scientific discoveries. The findings has tracked the impact of open government data programs, and they provide an initial explanation of how open government data are useful to a variety of scientific research communities of the benefits.

Li (2019) discussed that Text data has become one of the most common kinds of tourism big data as the Internet has grown in popularity. Text mining of such data has a lot of potential to inspire tourism practitioners as an efficient technique of expressing tourists' perspectives. Various text mining techniques have been presented and applied to tourism analysis in the last decade to construct tourism value analysis models, establish tourism recommendation systems, create tourist profiles, and make regulations for overseeing tourism markets. His paper has aimed to present a complete and up-to-date overview of text mining techniques that have been, or have the potential to be, applied to modern tourism text data sources, with an appreciation of the complexity owing to this broad range of approaches and tourism text data sources.

**Research Design and Methods**

The datasets used in this study were collected from three sources which include UAE open data portal (bayanat.ae), Statistica company, and UAE national news outfits. These datasets come in different file types such as MS Excel (xlsx), comma-separated-values (CSV), and web contents. Forty (40) tourism – related data files were culled from the UAE open data portal. These files comprise of 2015 and 2016 records. Five (5) data files were collected from Statistica portal which include datasets on UAE hotels' occupancy and tourism records for the years 2017, 2018, 2019, 2020, and 2021. Five(5) datasets were scraped from the UAE tourism websites covering the years 2020 and 2021. To ensure the accuracy of the data collected, the investigators randomly retrieved again the online files and information on different dates to check the data consistency.

Following the data collection, appropriate tools have been selected by the investigators. Several open-source data mining tools which are free software and downloadable from internet were evaluated. Two criteria were used to select the most appropriate software to use in processing and analyzing the tourism-related datasets which are Perceived Usefulness and Perceived Ease of Use. These criteria are based on Technology Acceptance Model (TAM). The TAM is an information systems theory that models how users come to accept and use a technology. Using the acceptability criteria for selection, the investigators agreed and decided to use the Pandas and RapidMiner tools. Both data mining software were used to process and analyze the collected datasets. Specifically, Pandas tool was used for data processing while Rapid Miner was used for data analysis and data mining.

The investigators of the study used the Big Data Intelligent Processing and Analysis model (Fig. 2) to process and analyze the multi-source data which were collected. This model is suitable for processing data from varying sources, huge data volume, and complex in nature. As data size is big and data type is diverse, big data technology is required for data processing and analysis (Guo et. al, 2015).

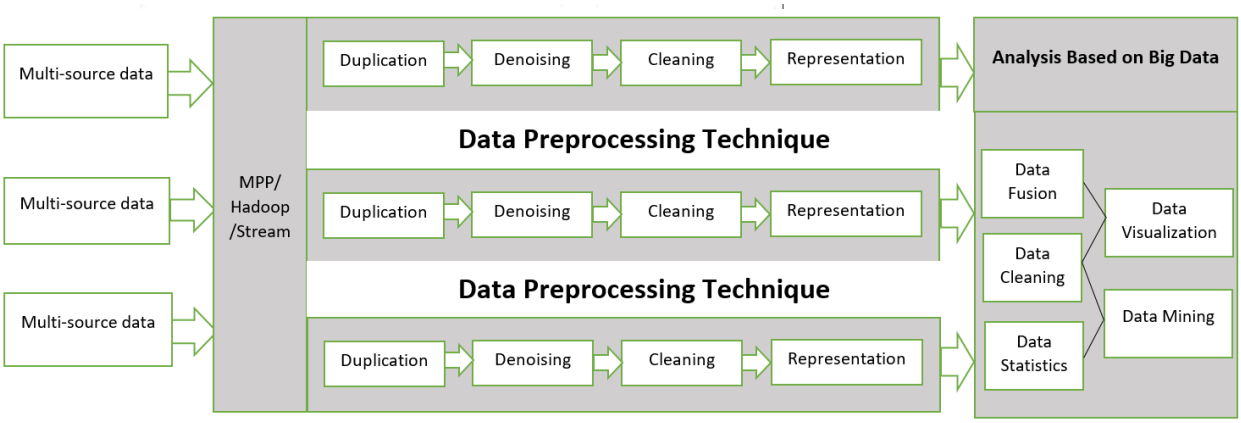


**Fig. 2 - Big Data Intelligent Processing and Analysis (NRCISP 2015)**

The model has three main phases namely – Data Sourcing (Multi- source data), Data Pre-processing (Data preprocessing technique), and Data Analysis (Analysis based on big data). In the data sourcing phase, the investigators of this study gathered all the tourism-related information such as hotels occupancy, apartments occupancy, and museums and parks visits. All the multi-source data were consolidated into three large spreadsheets which were categorized by year. This consolidation process involved manual such as selecting and classifying data types and automated tasks such as merging and juxtaposing related data.

Data pre-processing phase includes deduplication, de-noising, cleaning, and representation stages. The proponents utilized Pandas tool for data pre-processing because it has the capability to generate conditional queries for data. Data deduplication refers to a technique for eliminating redundant data in a dataset. In the process of deduplication, extra copies of the same data were deleted, leaving only one copy to be stored. Data were analyzed to identify duplicate byte patterns to ensure that there is only a single instance in every single file. Every copy of data was compared with other sheets and the duplicate data were discarded. Denoising is a technique used by the proponents that removed "noise" in the consolidated data. This was done to sharpen a fuzzy picture or aid in character recognition (e.g. differentiating a "6" from a "b" in text). In this study, there are two types of noise that were identified - label noise and attribute noise. Label noise occurs when a hotel type, room type, or nationality is incorrectly labeled. Label noise can be attributed to several causes, such as subjectivity during the labeling process, data entry errors, or inadequacy of the information used to label each item. On the other hand, the attribute noise refers to corruptions in the values of one or more attributes such as erroneous attribute values; missing or unknown attribute values; and incomplete attributes or "do not care" values.

Data cleaning reduced errors and improved the data quality. Errors were corrected in data and bad records were eliminated. Blank rows and unnecessary labels were deleted. Data cleaning was done to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. In this stage, the proponents followed the iterative two-step process consisting of discrepancy detection and data transformation. For missing values, during the process to seize data from all sorts of data sources, there were many cases when some fields are left blank or contain a null value. The last stage which is the Data Representation aimed to achieve substantial data reduction to a manageable size, while preserving important characteristics of the original data, and robustness to random noise. In this stage, the investigators carefully included

only the noise which may be useful for the analysis. Kaufmann (2011) propounded that while data noise are mainly discarded from the analyzed set because they do not comply with the general behavior or model of the data, some outliers should be considered as these may uncover interesting or critical patterns.

The data analysis phase involves five elements namely – data fusion, data clustering, data statistics, data visualization, and data mining. RapidMiner tool was utilized for these data analysis processes. In the data fusion, the preprocessed three consolidated tourism datasets were automatically re-consolidated into single master file. This stage produced the final single-copy of consolidated tourism dataset which was processed by RapidMiner tool to generate data clusters and statistics. Moreover, the investigators utilized the RapidMiner's feature to provide visualization of arbitrary models called Self-organizing Map (SOM). A SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space called a map. Additionally, RapidMiner generated data mining reports based on the clustering, pattern discovery, and classification techniques used in this study. These reports yielded the critical information that revealed the patterns, relationships, and forecasts of data. Finally, the results and reports generated by RapidMiner were carefully analyzed and interpreted by the investigators of this study.

## Presentation and Analysis of Results

This section presents the results of the data mining conducted by the investigators using the UAE open-data sources.

### Hotel StayingTourists Per Month (2015-2020)

Fig. 3 illustrates the average monthly number of tourists staying in hotel for the period 2015 until 2020. The top 24 countries visiting UAE are India, Saudi Arabia, UK, Russia, Oman, Pakistan, USA, China, France, Germany, Philippines, Kuwait, Nigeria, Italy, Kazakhstan, Sudan, Ukrain, Canada, Iran, Ethiopia, Jordan, Uzbekistan, and Bahrain. Other countries with minimal visit number are counted under Others in the chart. This finding is consistent with the figures released by the Dubai tourism statistics official website on the top visitors in the UAE (https://www.dubai-online.com/essential/tourism-statistics/).

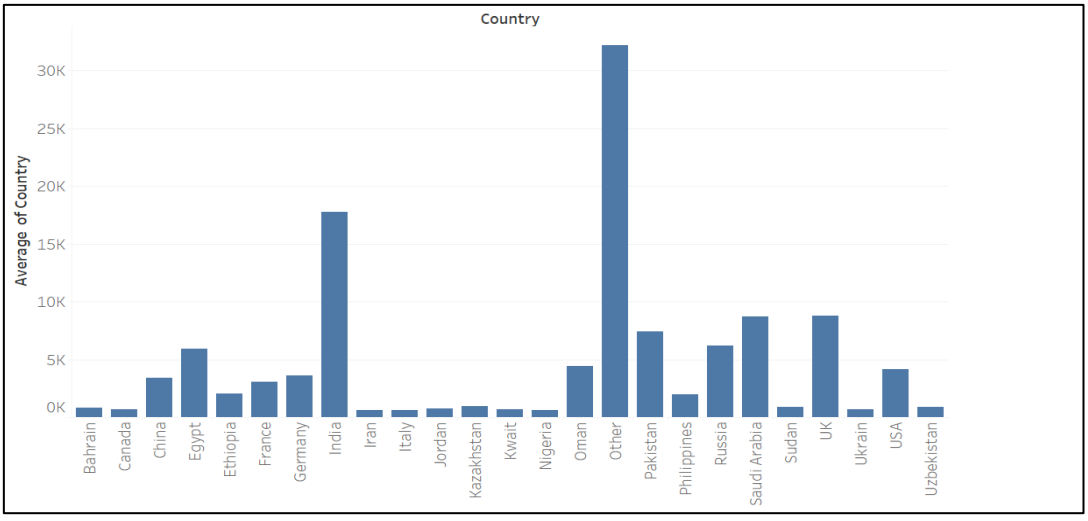


**Fig. 3 Average Number of Hotel Staying Tourists Per Month (2015-2020)**

Fig. 4 presents the monthly average number of tourists per nationality for the period 2015 until 2020. The chart reveals that December and January are the months which the UAE recieves the highest number of hotel-staying visitors. Conversely, August and November have the lowest average number of visitors. The August decline may be due to the warm weather condition of the country while the November figure may imply that visitors are delaying their visit to December as this is an eventful month with fair weather. The figure also reveals that UAE receives diverse nationalities during the year except for August and December. This trend may be attributed to the aformentioned warm weather condition during August and the tradition of many nationalities to visit their own families during December holidays.

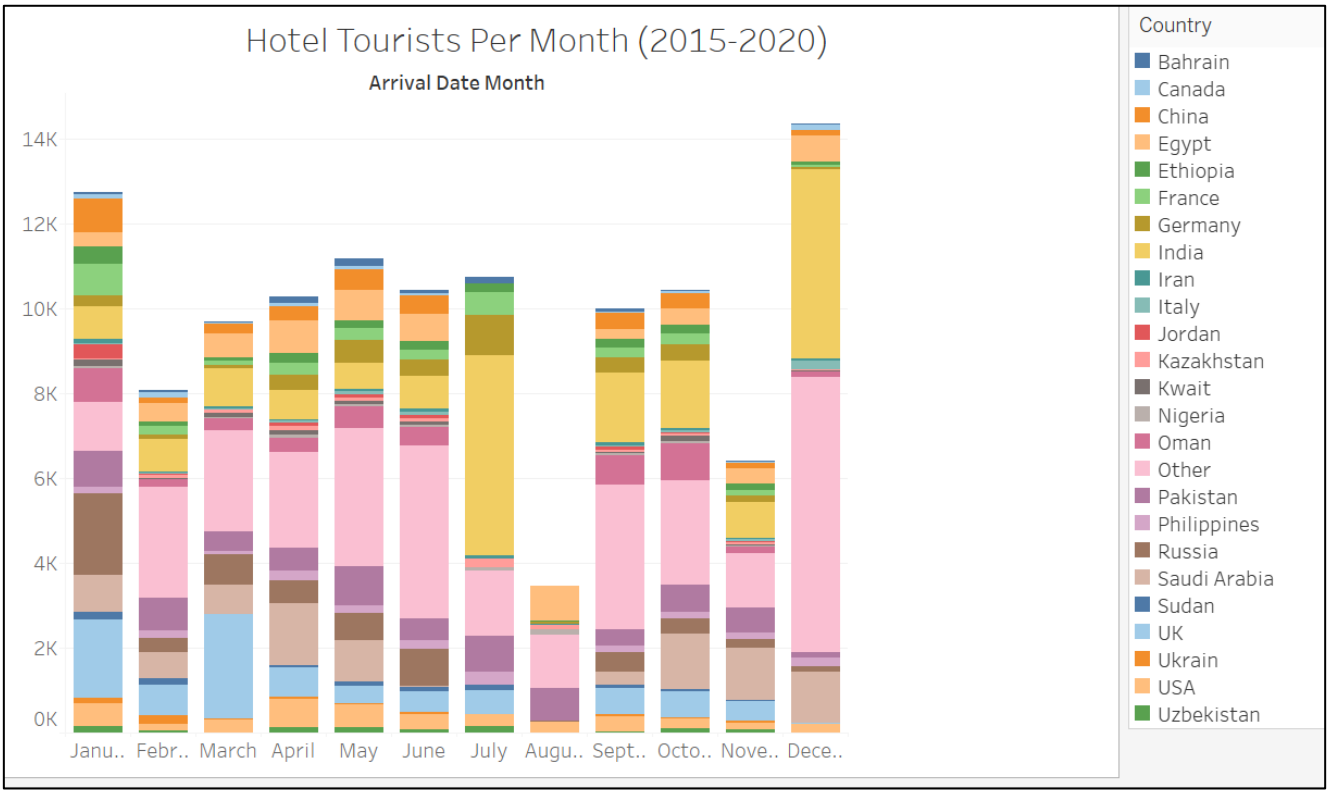


**Fig. 3 – Average Hotel Staying Tourists (2015-2020)**

Fig. 4 indicates the number of hotel-staying tourists for the period 2015 until 2020. The chart indicates that the number of tourists who stayed in City hotels is almost double the percent of tourist who stayed in Resort hotels. This implies that there is far more demand for city hotels than resort hotels. Though, there are other factors that may have caused this observation such as hotel prices, availability of rooms, among others. Interestingly, the chart also reveals that more Chinese prefer to stay in a resort hotel than city hotel while US, Egypt, Jordan, Germany, and Pakistan nationals choose otherwise.

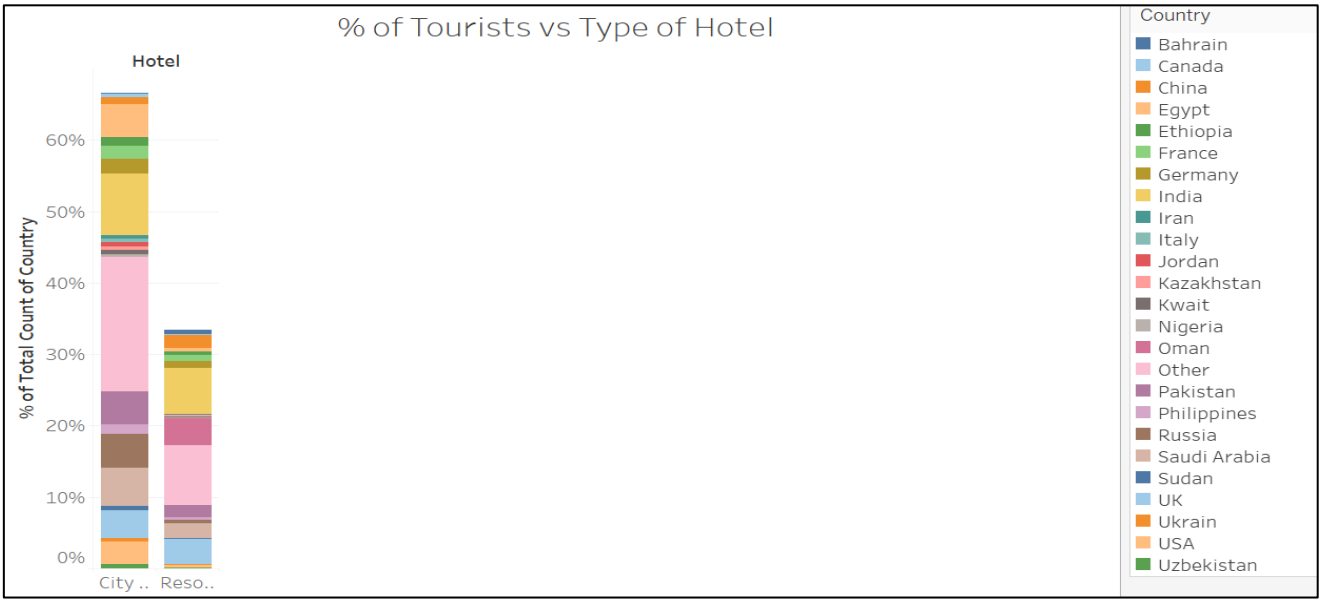


**Fig. 4 - % of Tourist Per Type of Hotel (2015-2020)**

Fig. 5 shows that tourists from Pakistan and Philippine stay longer in hotel as compared with other nationalities. This information may be useful for hotel management to plan for suitable or additional services for long-staying as well as short-staying tourists to meet their needs.

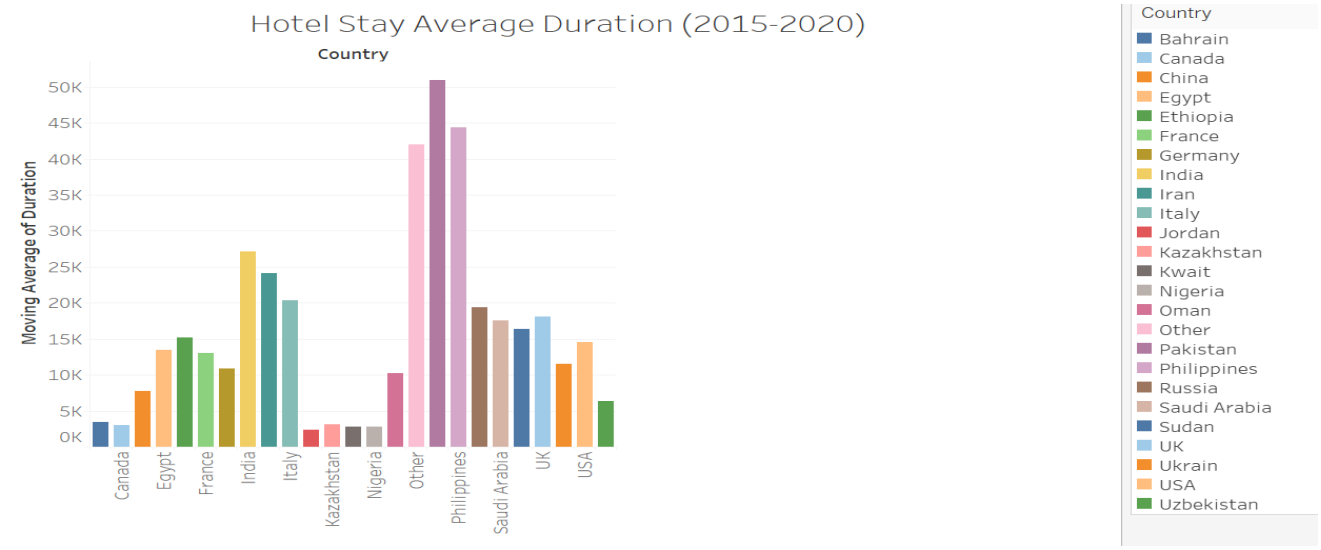


**Fig 5 – Hotel Stay Per Nationality**

Fig. 6 shows that 2-bed hotel rooms have low demand for the months of August and November. The demand for 1-bed and family hotel rooms both peak in December and January while the demand for 2-bed hotel room steeply peaks in June and December. Throughout the year, there is a good demand for 2-bed hotel rooms.

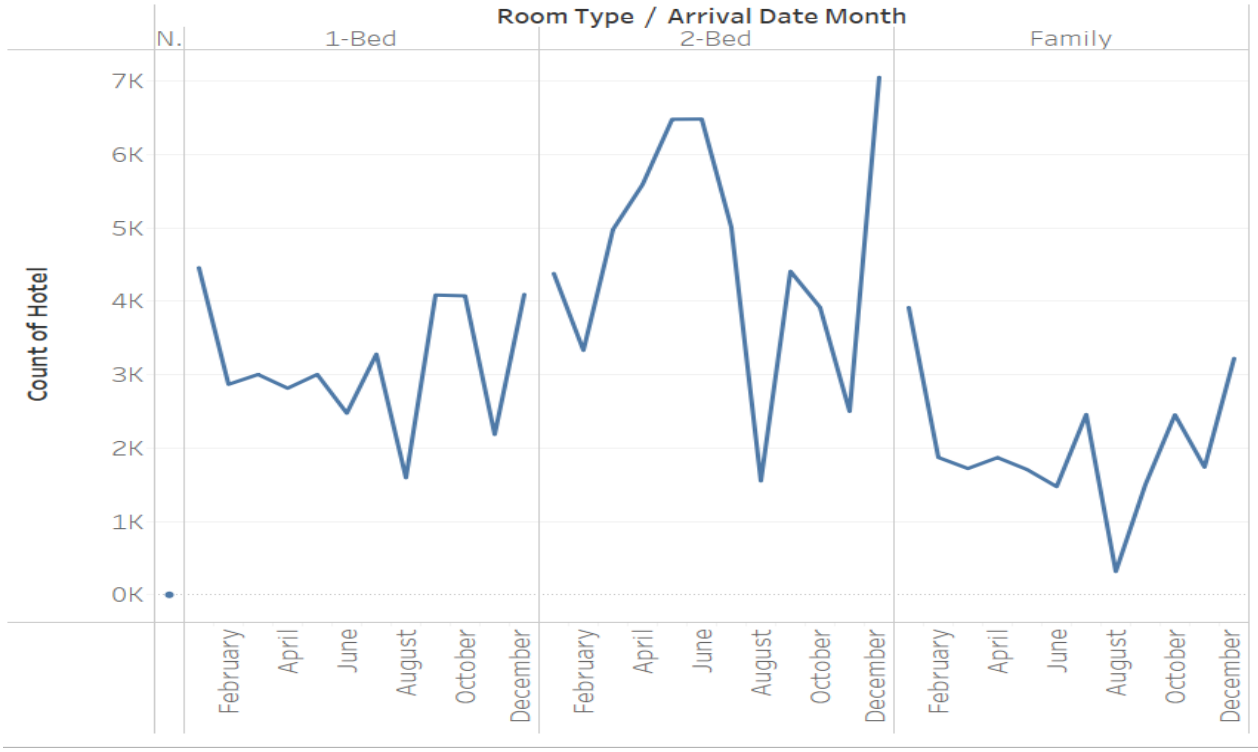


**Fig. 6 – Demand for Room Type Per Month (2015-2020)**

Fig 7 presents the comparison of the demand for room type vis-à-vis hotel type. The chart reveals that there is a fair and equal demand for 1-bed, 2-bed, and family rooms in city hotels. Whereas, the demand for 2-bed and family room is significantly higher than the demand for 1-bed room in resort hotels.

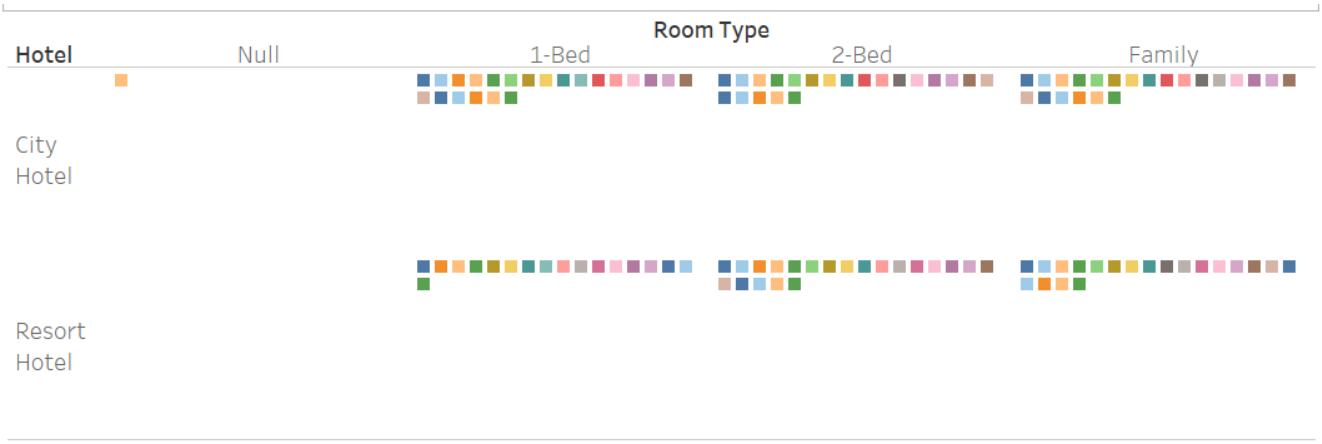


**Fig 7 – Demand for Hotel and Room Types (2015-2020)**

**Trend and Forecasts of Hotel Staying Tourists**

Fig. 8 shows the trend of number of hotel-staying tourists for the period 2015 until 2021. The arrival data of tourists for the year 2021 was not included in the previous presentation due to the incomplete data attributes needed for processing. The chart illustrates the steady increase of hotel-staying tourists for the period 2015 until 2019 in the UAE until its sharp fall in year 2020 when pandemic peaked. However, there is an increase in the number of arrivals in year 2021 which can be attributed to the opening of several countries as they start to open the tourism

industry. This trend also implies that UAE is one of the resilient countries who managed well the pandemic and it is back on its feet being a top tourist destination for the past several years.

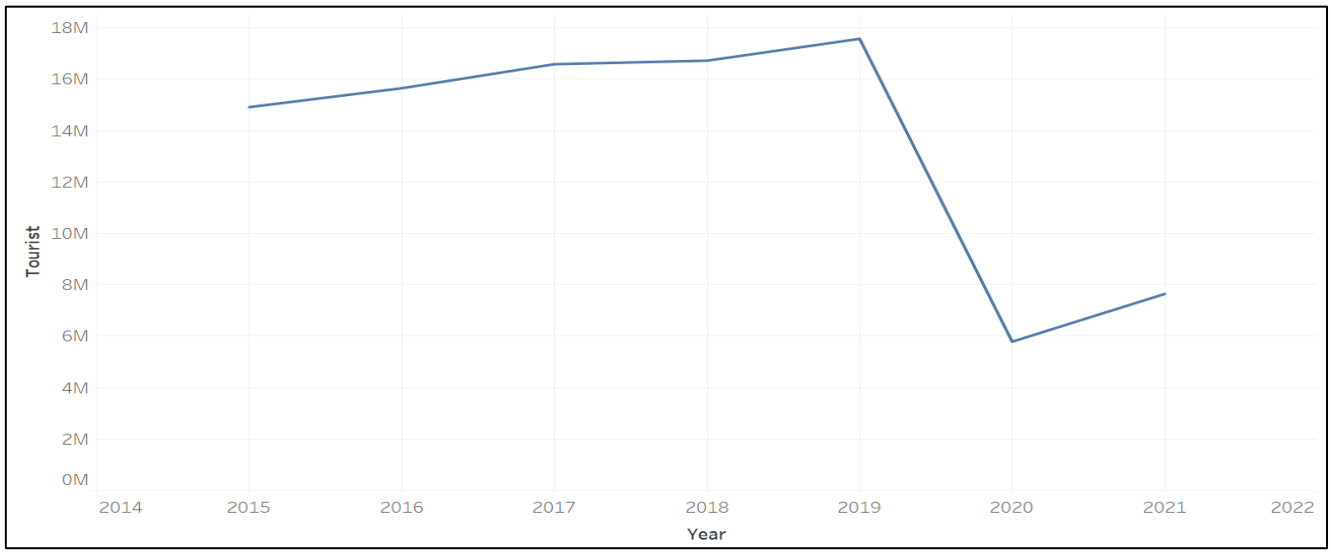


**Fig 8 – Trend of Hotel Staying Tourists Arrival (2015-2021)**

Fig. 9 and Fig. 10 illustrate the 2022 and 2023 period forecasts based on the 2015-2021 hotel-staying tourists' data. Using the exponential smoothing model, UAE will have around 11% and 10% increase of number of tourists from 2021 and 2022 respectively. The forecasted 11% increase is much lower than the estimate of the Tourism sector which is at least 30% based on its first quarter tourists' arrival data. However, the first quarter data include the Dubai EXPO2020, a big tourist attraction, which may be a cause a sudden surge of tourists only for this seasonal event.

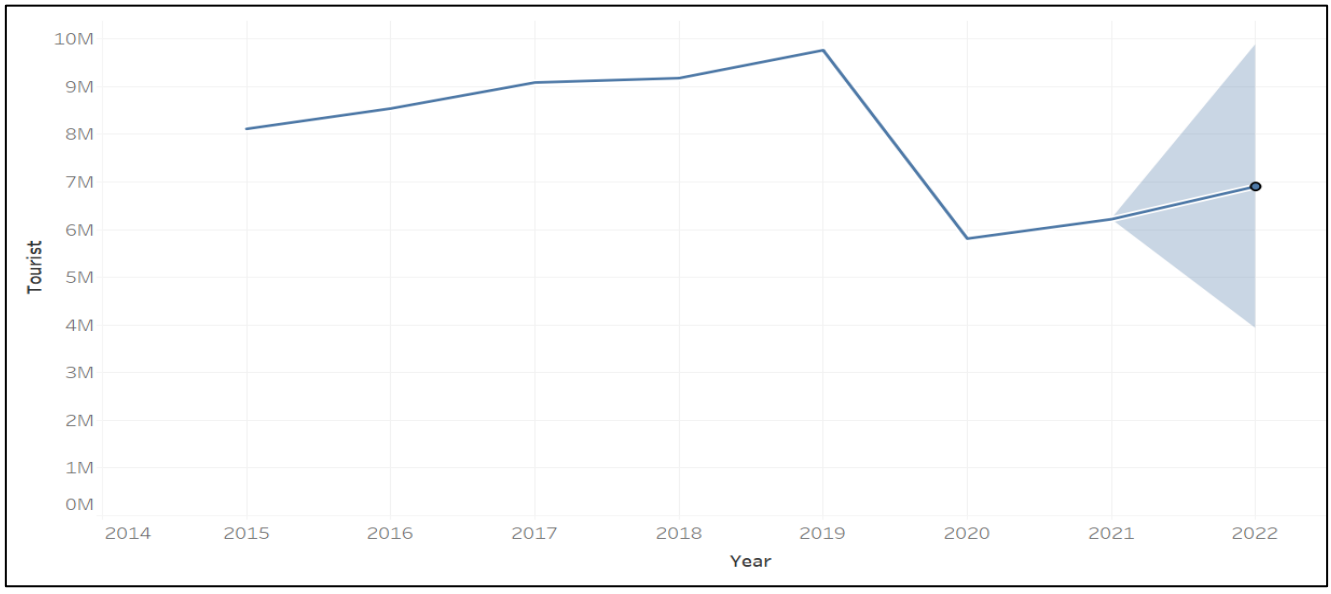


**Fig. 9 – Hotel Staying Tourists Forecast 2022**

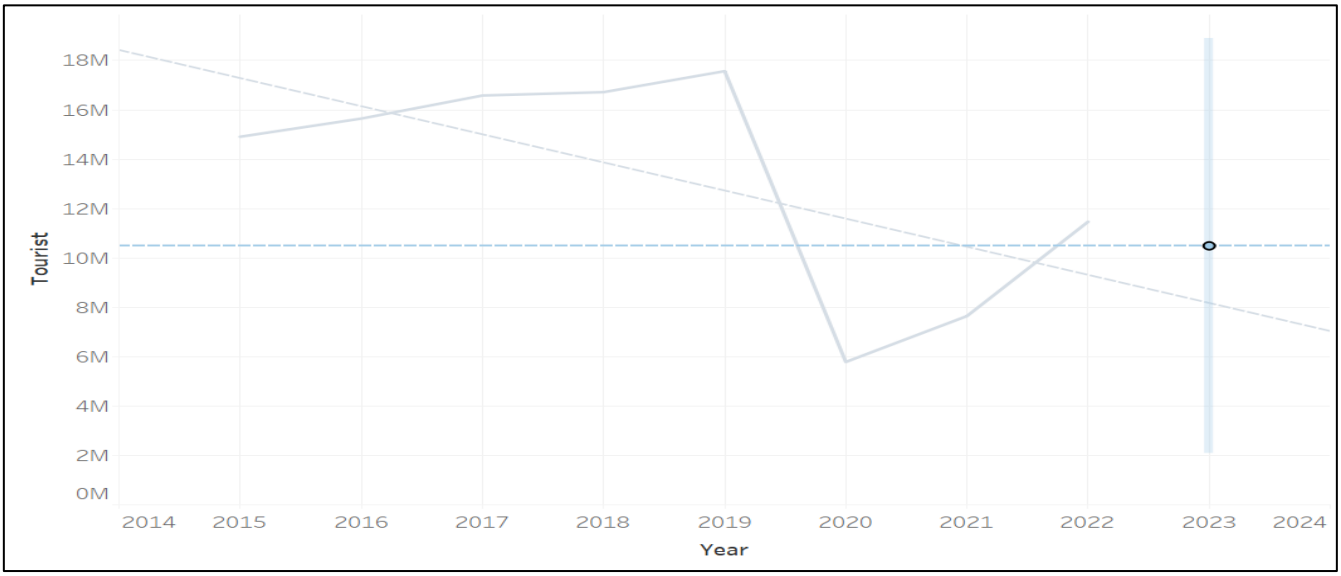


**Fig. 10 – Hotel Staying Tourists Forecast 2023**

**Other Key Findings**

Using data partitioning and decision tree to process the consolidated tourism data, the following knowledge were also discovered:

- GCC (Oman, Saudi Arabia, Bahrain, and Kuwait) tourists most likely select 2-bed or family type of room.
- Non-GCC tourists most likely visit UAE during the last quarter of the year.
- GCC tourists most likely visit UAE during the first quarter of the year.
- Western (North and South America, Europe, Australia and New Zeeland) tourists have diverse preference for City Hotel or Resort Hotel.

**Conclusion**

1. Data mining can undoubtedly play an important role in travel and tourism. It is necessary to determine the data mining techniques to use to appropriately analyze the outcomes. To generate relevant information insights on tourism business, regression and classification data mining techniques are used.
2. Using the UAE open-data sources, tourism industry were severely impacted by the pandemic in the face of harsh travel restrictions and restricted borders. This study was limited to pre-pandemic information due to the lack of adequate post-pandemic data. As a result, there are sufficient chances for additional study aimed at applying the findings to post-pandemic market situations.

**Recommendations:**

1. Data mining applications require a significant investment of resources from the government creating them.
2. Future research could examine the application of data analysis and compare results from different attractions and tourist sites around the world.

3. Government should ensure the safety of the travellers and should be in a controlled environment upon their arrival to the country.
4. The travel and tourism industry, as well as businesses, must recognize the potential benefits and utility of data.
5. The user must be skilled in both the domain area of application and the data mining technology and tools in order to apply data mining successfully. Domain expertise is required to discover acceptable business concerns for data mining development.

## References


[1] Lau K.N. et al. (2001) Cornell Hotel and Restaurant Administration Quarterly, 42(6), 55-62

[2] Leifeng, et al. (2015). A Study of the Application of Big Data in a Rural Comprehensive Information Service. ResearchGate Publication. https://www.researchgate.net/figure/Big-data-intelligent-processing-and-analysis_fig4_277943364

[3] Manoj et al. (2012). Data Mining for Travels and Tourism. ResearchGate Publication. https://www.researchgate.net/publication/236966138

[4] Wang, J (2003). Data Mining: Opportunities and Challenges. Idea Group Publishing

[5] Guo et.al. (2015). A Study of the Application of Big Data in a Rural Comprehensive Information Service.

[6] Kaufmann(2011). Data Mining: Concepts and Techniques. Elsevier Inc. Publication

[7] Yao (2003)., A step towards the foundations of data mining, Proceedings of the SPIE: Data Mining and Knowledge Discovery: Theory, Tools, and Technology, 5098, 254-263

[8] Sowa (1984). Conceptual Structures, Information Processing in Mind and Machine Addison-Wesley, Reading, Massachusetts, 1984Addison-Wesley, Reading, Massachusetts, 1984

[9] chine, Addison-Wesley, Reading, Massachusetts, 1984

[10] Lin (2005). Ohsuga, S., Liau, C.J. and Hu, X. (Eds.),Foundations and Novel Approaches in Data Mining, 75-97, Springer-Verlag

[11] Minelli, et al. (2013). Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses

[12] Gottfried A, Hartmann C, Yates D. Mining Open Government Data for Business Intelligence Using Data Visualization: A Two-Industry Case Study. Journal of Theoretical and Applied Electronic Commerce Research. 2021; 16(4):1042-1065. https://doi.org/10.3390/jtaer16040059

[13] Yan, A., & Weber, N. (2018, March). Mining open government data used in scientific research. In International Conference on Information (pp. 303-313). Springer, Cham.

[14] Lausch, A., & Schmidt, A., et al (2015, January). Data mining and linked open data – New perspectives for data analysis in environmental research; Ecological Modelling: Vol 295; (pp 5-17)

[15] Li Q, Li S, Zhang S, Hu J, Hu J. A Review of Text Corpus-Based Tourism Big Data Mining. Applied Sciences. 2019; 9(16):3300. https://doi.org/10.3390/app9163300