

CUSTOMER SEGMENTATION MODELLING AND ANALYTICS

MINOR PROJECT II

Submitted by:

Aayushie Prasad (19104008)

Saransh Gupta (19104015)

Siddharth Khandelwal (19104058)

Under the supervision of

Dr Neetu Sardana



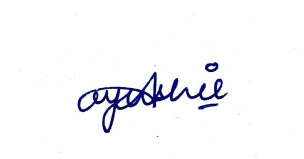
Department of CSE/IT

Jaypee Institute of Information Technology, Noida

MAY 2022

ACKNOWLEDGEMENT

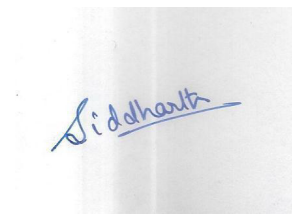
We would like to express our sincere gratitude to Dr Neetu Sardana, Associate Professor, Department of CSE/IT, Jaypee Institute of Information Technology, Noida for her generous guidance, continuous encouragement, help and valuable suggestions throughout the present work.



Aayushie Prasad (19104008)



Saransh Gupta (19104015)



Siddharth Khandelwal (19104058)

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgement has been made in the text.

Place: JIIT, Noida

Date: 17 May 2022

Name: Aayushie Prasad

Enrolment No.: 19104008

Name: Saransh Gupta

Enrolment No.:19104015

Name: Siddharth Khandelwal

Enrolment No.:19104058

CERTIFICATE

This is to certify that the work titled “Customer Segmentation Modelling and Analytics” submitted by Aayushie Prasad, Saransh Gupta, Siddharth Khandelwal of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.



Digital Signature of Supervisor

Name of Supervisor: Dr Neetu Sardana

Designation: Associate Professor

Date: 15 May 2022

ABSTRACT

Hotel Industry records customer actions which generate a huge amount of data that can lead to crucial insights into customer behaviour and demands. This project aims to explore K-means clustering and Expectation-Maximisation clustering and compare the two models. For further modelling, the RFM analysis model is applied to the Hotel Customer Dataset. The original dataset contains 83,590 instances (customers) and 31 features. It comprehends three full years (2015-2018) of customer behavioural data. In addition, the dataset also contains demographic and geographical information about the customers. We perform dimensionality reduction using PCA (Principle Component Analysis) to filter out irrelevant features. Clustering is done using both algorithms to form customer segments having similar characteristics. Different internal measures for validation were performed to evaluate both the clustering processes. After evaluation, the clusters of the better clustering algorithm are integrated with the RFM model categories and customers belonging to different clusters are segmented based on RFM categorization. Vivid and detailed visualizations of different clusters are shown using the Microsoft Power BI tool.

Table of Contents

Page No.

<i>Abstract</i>	<i>i</i>
<i>Acknowledgement</i>	<i>ii</i>
<i>List of Tables</i>	<i>iv</i>
<i>List of Figures</i>	<i>v</i>
Chapter 1: INTRODUCTION	1
Chapter 2: LITERATURE REVIEW	
2.1 BACKGROUND STUDY	2
2.2 REQUIREMENT ANALYSIS	3
2.3 DETAILED DESIGN	4-7
Chapter 3: IMPLEMENTATION AND ANALYSIS	
3.1 IMPLEMENTATION	8-10
3.2 EXPERIMENTAL RESULTS AND ANALYSIS	11-16
3.3 CONCLUSION AND FUTURE SCOPE	17
References	18

List of Tables

Table	Title	Page
2.1	Literature Review Of Research Papers	2
2.2	Dataset Specification	4
2.3	Detailed Description of Dataset	4-6

List of Figures

Figure	Title	Page
2.1	Minor 2 Project Workflow	7
3.1	Count of Customers Monthwise	11
3.1	Customers visiting per year	11
3.2	Rooms Booking per person.....	11
3.2	Number of Persons per Night.....	12
3.4	Percentage of Amenities	12
3.5	Data points per Cluster using k means	12
3.6	Clusters per Age	13
3.7	Clusters per RFM Segment	13

INTRODUCTION

Customers have a variety of demands and want, as the hotel industry is well aware. Hotel chains have utilised a variety of segmentation criteria and approaches to better identify and understand client groups and supply them with more appropriate products and services to meet their various needs and standards. Segmentation is also necessary so that the organisation may build profitable segments and respond to them based on their competitive advantages. However, many businesses struggle to identify the optimal consumer segments for marketing campaigns and advertising strategies. This results in ineffective loyalty programmes and promotions, as well as a waste of marketing initiatives and advertising strategies. Customer retention has become a priority for marketing executives.

Client loyalty programmes and loyalty cards have attracted the attention of marketing researchers and analysts as a vital strategy for improving customer retention rates. Customers are segmented into several status categories by loyalty programmes based on their cumulative spending. Consumer loyalty, on the other hand, encompasses more than just payment considerations. Customer loyalty has attitudinal as well as behavioural components. Customers' spending behaviour and the frequency of their visits is being referred to as behavioural loyalty. While, major components of attitudinal loyalty include satisfaction, commitment, and trust.

BACKGROUND STUDY

sno	Paper ID	Objective	Techniques	Evaluation Metrics	Outcome	Critical Analysis/Gaps																																							
1	Al Khayrat et al. J Big Data (2020)	A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA	Dimensionality reduction of the dataset using PCA and Autoencoder neural network (ANN). K-means clustering was applied to both reduced and original datasets. Finally, evaluated how the reduction method impacts the clustering task.	Internal cluster validation: 1) <i>Silhouette coefficient</i> : measures how well the features are clustered. Ranged between – 1 and + 1, the silhouette indicates that the object is well matched to its cluster and poorly matched to neighbouring clusters and vice versa. 2) <i>DB index The Davies–Bouldin index (DBI)</i> : average measure of similarity of each cluster with its most similar cluster where similarity is the ratio within-cluster distances to between-cluster distances. The lower the value of the DB index, the better the clustering is.	Silhouette scores after K-means clustering: 0.682 with 3 clusters for ANN, 0.581 on the original dataset, and 0.476 for 2 clusters with PCA.	In future work, we can Tune and Optimize Autoencoder using PCA principles by building custom constraints for our Autoencoder for tuning and optimization. So the incorporation of the PCA properties will bring significant benefits to an Autoencoder, such as resolving overfitting via regularization.																																							
2	International Journal of Contemporary Economics and Administrative Sciences ISSN: 1925-4423 Year:2018	Customer Segmentation By Using RFM Model And Clustering Methods: A Case Study In Retail Industry	Three major steps. Step 1) Data cleaning and Transformation. Step 2) Data analyzed using RFM analysis, two-step cluster analysis and K-means clustering. Step 3) Results are analysed and presented.	A log-likelihood method is chosen to measure the distance and BIC)as clustering criteria. The 3 clusters are evaluated by the results of a two-step cluster analysis. <i>No clear data on proposed model 2.</i>	<i>The two-step clustering method</i> : There is a 60% discrepancy between the first and the current model. <i>No clear outcome of model 2.</i>	BIC (Schwarz's Bayesian Criterion) penalizes a model more severely for the number of parameters compared to AIC (Akaike Information Criteria). (both are model selection criteria)																																							
3	A.J. Christy et al. / Journal of King Saud University – Computer and Information Sciences 33 (2021) 1251–1257	RFM ranking – An effective approach to customer segmentation	RFM is performed on online retail transactional data. Then K-means and Fuzzy C-means algorithms are applied to form clusters. A novel approach to choosing centroid in K-means is used. The results obtained from the methodologies are compared with one another by their iterations, cluster compactness and execution time.	<i>Comparative Analysis of RM K-Means with others</i> : K-Means, Fuzzy C-Means, RM K-Means respectively. <i>Iterations</i> : 4, 193, 2 <i>Time Taken (in seconds)</i> : 2.0035, 24.7988, 1.4917 <i>Average silhouette width</i> : 0.38, 0.06, 0.49	It is observed that the proposed RM K-Means consumes lesser time than the other two techniques because of the lesser number of iterations. The average silhouette width of RM K-Means is greater than that of Fuzzy C-Means clustering and the K – Means clustering.	Future work includes studying the performance of the customers in each segment such as the products which are bought frequently by the members of each segment. This would help better in providing better promotional offers to specific products.																																							
4	2010 IEEE International Conference on Data Mining Workshops	Using SOM-Ward clustering and predictive analytics for conducting customer segmentation	Step 1) SOMWard Clustering is used to segment the customer base according to their spending amount, demographic and behavioural characteristics. Step 2) Three classification models - the support vector machine (SVM), the neural network, and the decision tree, are employed to classify high-spending and low-spending customers. Step 3) The performance of the three classification models is evaluated and compared.	Comparison of 3 models:[1] <table><tr><th>Classification models</th><th>Accuracy (%) (Re-substitution)</th><th>Accuracy (%) (Cross-validation)</th><th>Area under the ROC curve</th></tr><tr><td>Support vector machine</td><td>81.25</td><td>80.79</td><td>0.89</td></tr><tr><td>Neural network</td><td>80.31</td><td>79.60</td><td>0.87</td></tr><tr><td>Boosted decision tree</td><td>82.49</td><td>80.58</td><td>0.88</td></tr><tr><td>Decision tree without boosting</td><td>81.73</td><td>80</td><td>0.86</td></tr></table> Comparison of the three ensemble methods. [2] <table><tr><th rowspan="2">Ensemble method</th><th colspan="3">Training data</th></tr><tr><th>Overall accuracy (%)</th><th>Accuracy of predicting high-spending customers (%)</th><th>Accuracy of predicting low-spending customers (%)</th></tr><tr><td>Voting</td><td>82.37</td><td>84.83</td><td>79.92</td></tr><tr><td>Confidence-weighted voting</td><td>82.39</td><td>84.22</td><td>80.55</td></tr><tr><td>Highest confidence win</td><td>82.29</td><td>83.46</td><td>81.11</td></tr></table>	Classification models	Accuracy (%) (Re-substitution)	Accuracy (%) (Cross-validation)	Area under the ROC curve	Support vector machine	81.25	80.79	0.89	Neural network	80.31	79.60	0.87	Boosted decision tree	82.49	80.58	0.88	Decision tree without boosting	81.73	80	0.86	Ensemble method	Training data			Overall accuracy (%)	Accuracy of predicting high-spending customers (%)	Accuracy of predicting low-spending customers (%)	Voting	82.37	84.83	79.92	Confidence-weighted voting	82.39	84.22	80.55	Highest confidence win	82.29	83.46	81.11	The results of the analysis demonstrate that the combined method of SOM-Ward clustering and predictive analytics can potentially be effective in conducting customer segmentation. The results of market segmentation, incorporated with the results of predictive analytics, are more prospective and predictive.	<i>SOM-Ward clustering</i> may warrant a good understanding of the current customer base but they cannot necessarily provide information about the potential value of a segment. <i>Predictive Analytics</i> : The starting point of predictive analytics inevitably requires more a priori knowledge than unsupervised learning does, making the knowledge gained in it less significant than in SOM-Ward model.
Classification models	Accuracy (%) (Re-substitution)	Accuracy (%) (Cross-validation)	Area under the ROC curve																																										
Support vector machine	81.25	80.79	0.89																																										
Neural network	80.31	79.60	0.87																																										
Boosted decision tree	82.49	80.58	0.88																																										
Decision tree without boosting	81.73	80	0.86																																										
Ensemble method	Training data																																												
	Overall accuracy (%)	Accuracy of predicting high-spending customers (%)	Accuracy of predicting low-spending customers (%)																																										
Voting	82.37	84.83	79.92																																										
Confidence-weighted voting	82.39	84.22	80.55																																										
Highest confidence win	82.29	83.46	81.11																																										

Table 2.1: Literature Review Of Research Papers

REQUIREMENT ANALYSIS

Language Used - Python because it is currently the most important programming language for Machine Learning. Also, humans are able to interpret it easily which makes it easier to build models for machine learning.

Libraries Used:

- **Pandas** - It is used to import data and create python objects with rows and columns and can also be used to write data into the file.
- **NumPy** - It contains multi-dimensional arrays such as matrix data structures. It is used to perform statistical and algebraic mathematical operations on arrays.
- **Matplotlib** - It is a visualization library. Used to create interactive graphs and charts.
- **Scikit Learn** - It provides us with tools which make the implementation of machine learning in python a lot easier. It helps in implementing supervised and unsupervised algorithms by just importing their libraries.
- **Seaborn** - It is again a data visualization library based on **matplotlib** used for making a high-level interface for drawing statistical graphs.
- **Yellowbrick** - Yellowbrick is a suite of visualization and diagnostic tools that will enable quicker model selection. It's a Python package that combines scikit-learn and matplotlib.
- **Klib** - klib is a Python library for importing, cleaning, analyzing and preprocessing data.

DETAILED DESIGN

DATASET USED

Specifications Table

<i>Subject</i>	Tourism, Leisure and Hospitality Management
<i>Specific subject area</i>	Marketing, Customer Relationship Management, and Revenue Management
<i>How data were acquired</i>	Extraction from the hotel Property Management System (PMS) SQL database
<i>Data format</i>	Mixed (raw and pre-processed)
<i>Parameters for data collection</i>	The unit of analysis of the dataset is a customer. A full three-year period of data was collected (2015 to 2018). All personal related data were transformed or anonymized to guarantee the privacy and prevent the hotel or guests' identification. Time-related variables were accounted for based on the last day of the extraction period. The last day of the extraction period is December 31, 2018.
<i>Description of data collection</i>	Data was extracted via TSQL queries executed in the production server, using Microsoft SQL Studio Manager. Python was employed to perform summary statistics.
<i>Data source location</i>	The data came from a four-star hotel located in Lisbon, Portugal, Europe. In Portugal, hotels' star classification scale varies from 1 to 5, with one-star being the low-end quality hotels and five-star being the high-end quality hotels.
<i>Data accessibility</i>	Data is available from http://dx.doi.org/10.17632/j83f5fsh6c.1

Table 2.2: Dataset Specification

Data Description

HotelCustomersDataset.tsv is provided in tab-separated value format. The unit of analysis of the dataset is the customer, each instance (row) of the dataset represents one customer. Each variable (column) represents a characteristic or description of the customer.

Variable	Type	Description
<i>ID</i>	Numeric	Customer ID
<i>Nationality</i>	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [1]
<i>Age</i>	Numeric	Customer's age (in years) at the last day of the extraction period.
<i>DaysSinceCreation</i>	Number	Number of days since the customer record was created (number of days elapsed between the creation date and the last day of the extraction period)

<i>NameHash</i>	Categorical	Name of the customer's SHA2–256 hash string. A hash string is a string resulting from a mathematical function that maps a string of arbitrary length to a fixed-length [2].
<i>DocIDHash</i>	Categorical	SHA2–256 hash-string of the identification document number the customer provided at check-in (passport number, national ID card number, or other)
<i>AverageLeadTime</i>	Numeric	The average number of days elapsed between the customer's booking date and arrival date. In other words, this variable is calculated by dividing the sum of the number of days elapsed between the moment each booking was made and its arrival date, by the total of bookings made by the customer
<i>LodgingRevenue</i>	Numeric	Total amount spent on lodging expenses by the customer (in Euros). This value includes room, crib, and other related lodging expenses
<i>OtherRevenue</i>	Numeric	Total amount spent on other expenses by the customer (in Euros). This value includes food, beverage, spa, and other expenses
<i>BookingsCanceled</i>	Numeric	Number of bookings the customer made but subsequently cancelled (the customer informed the hotel he/she would not come to stay)
<i>BookingsNoShowed</i>	Numeric	Number of bookings the customer made but subsequently made a “no-show” (did not cancel but did not check-in to stay at the hotel)
<i>BookingsCheckedIn</i>	Numeric	Number of bookings the customer made, and which end up with a staying
<i>PersonsNights</i>	Numeric	The total number of persons/nights that the costumer stayed at the hotel. Person/nights of each booking is the result of the multiplication of the number of staying nights by the sum of adults and children
<i>RoomNights</i>	Numeric	Total of room/nights the customer stayed at the hotel (checked-in bookings). Room/nights are the multiplication of the number of rooms of each booking by the number of nights of the booking
<i>DaysSinceLastStay</i>	Numeric	The number of days elapsed between the last day of the extraction and the customer's last arrival date. A value of –1 indicates the customer never stayed at the hotel
<i>DaysSinceFirstStay</i>	Numeric	The number of days elapsed between the last day of the extraction and the customer's first arrival date (of a checked-in booking). A value of –1 indicates the customer never stayed at the hotel

<i>DistributionChannel</i>	Categorical	Distribution channel usually used by the customer to make bookings at the hotel
<i>MarketSegment</i>	Categorical	Current market segment of the customer
<i>SRHighFloor</i>	Boolean	Indication if the customer usually asks for a room on a higher floor (0: No, 1: Yes)
<i>SRLowFloor</i>	Boolean	Indication if the customer usually asks for a room on a lower floor (0: No, 1: Yes)
<i>SRAccessibleRoom</i>	Boolean	Indication if the customer usually asks for an accessible room (0: No, 1: Yes)
<i>SRMediumFloor</i>	Boolean	Indication if the customer usually asks for a room on a middle floor (0: No, 1: Yes)
<i>SRBathtub</i>	Boolean	Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes)
<i>SRShower</i>	Boolean	Indication if the customer usually asks for a room with a shower (0: No, 1: Yes)
<i>SRCrib</i>	Boolean	Indication if the customer usually asks for a crib (0: No, 1: Yes)
<i>SRKingSizeBed</i>	Boolean	Indication if the customer usually asks for a room with a king-size bed (0: No, 1: Yes)
<i>SRTwinBed</i>	Boolean	Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes)
<i>SRNearElevator</i>	Boolean	Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes)
<i>SRAwayFromElevator</i>	Boolean	Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes)
<i>SRNoAlcoholInMiniBar</i>	Boolean	Indication if the customer usually asks for a room with no alcohol in the mini-bar (0: No, 1: Yes)
<i>SRQuietRoom</i>	Boolean	Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes)

Table 2.3: Detailed Description of Dataset

<https://www.sciencedirect.com/science/article/pii/S2352340920314645>

Workflow/flow of operation

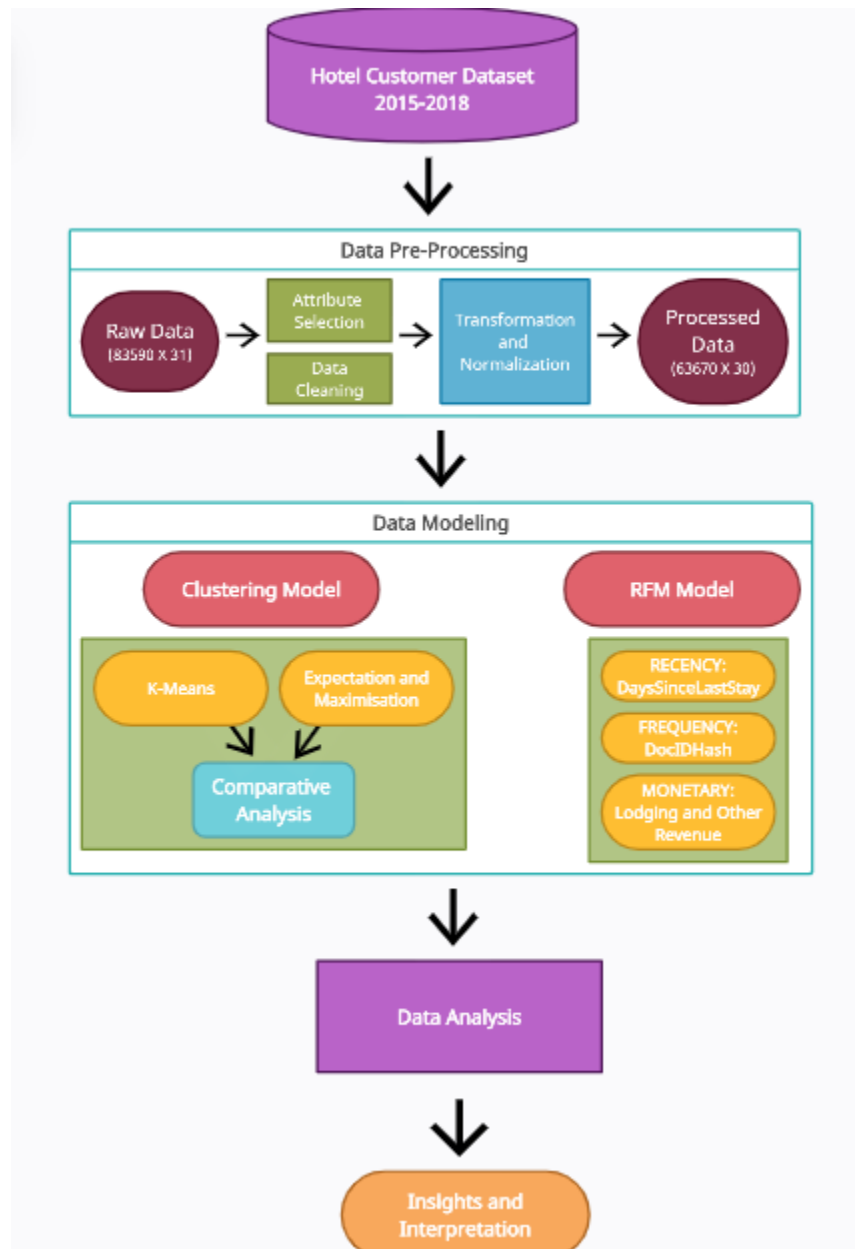


Fig.2.1: Minor 2 Project Workflow

IMPLEMENTATION

The proposed methodology of our project includes three major steps.

The first phase was related to pre-analysis efforts which refer to data cleaning and transformation.

- Null value imputation (only the Age variable has 4% null variables, it's been imputed using mode.)

- **Data cleaning:**

In our dataset, around 19920 instances have no information on all variables. Removing those instances does not affect the model, so those instances are removed.

- **Feature Transformation:**

- **Nationality:**

Nationality variables are having 188 unique categories (countries codes), to reduce the levels the countries are converted into 7 continents.

- **Amenities:**

Boolean variables like SRHighFloor, SRLowFloor, SRAccessibleRoom, SRMediumFloor, SRBathtub, SRShower, SRCrib, SRKingSizeBed, SRTwinBed, SRNearElevator, SRAwayFromElevato, SRNoAlcoholInMiniBar, SRQuietRoom represents the preference of the customer, combining all columns implies whether the customer has opted for any special requests and gives much more information than the variables. The Boolean variables are combined into a single variable named Amenities.

In the Second Phase, Data modelling using Clustering techniques such as K-means Clustering and Expectation-Maximization Clustering, and RFM modelling was performed on the processed dataset.

- **K-Means Clustering:** K-Means is a standard algorithm which takes the features and the number of clusters as inputs and partitions the data into the defined number of clusters such that the intra-cluster similarity is high. K-means algorithm was run for different number of clusters (0 to 10 clusters) initially on the processed dataset on the basis of attributes ('Age', 'DaysSinceCreation', 'AverageLeadTime', 'LodgingRevenue', 'OtherRevenue', 'PersonsNights', 'RoomNights', 'DaysSinceLastStay', 'DaysSinceFirstStay').

For each value of K, we calculated WCSS (Within-Cluster Sum of Square) and plotted the WCSS with the K value in order to find the optimal number of clusters (called the Elbow Method).

The optimal number of clusters = 4 using the Elbow Method.

Next, dimensionality reduction is applied using PCA to find a low-dimensional representation of the observation that explains a good fraction of the variance.

The optimal number of components determined =7. Thus using PCA before K-means clustering reduces dimensions and hence decreases computation cost by filtering out irrelevant features. After PCA analysis, the K-means algorithm is applied to get more accurate and defined clusters.

No. of Data points per Cluster after K-means:

Cluster 0: 8053 **Cluster 1:** 11804 **Cluster 2:** 21722 **Cluster 3:** 22091

- **EM Clustering:** The EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. Similar approach of applying PCA analysis and then EM clustering based on attributes('Age', 'DaysSinceCreation', 'AverageLeadTime', 'LodgingRevenue', 'OtherRevenue','PersonsNights', 'RoomNights', 'DaysSinceLastStay', 'DaysSinceFirstStay') is proposed on the processed dataset, resulting in 4 clusters of different sizes.

Comparative Analysis of Clustering Models

All the clustering techniques were evaluated using internal indices like the Silhouette Score/Coefficient.

Silhouette Score: Performance metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. Greater the silhouette coefficient of a clustering technique, the better the clustering model used for segmentation.

Silhouette Score for K-Means: 0.229

Silhouette Score for Expectation and Maximization: 0.074

Thus, the K-Means algorithm is best suited for Data Modelling through the clustering technique.

- **RFM Modelling:** The RFM segmentation is a solution to split our customers into segments(clusters) according to their recency, frequency and monetary values.

Applying RFM segmentation on the processed dataset.

Recency (R): How recently a customer has visited the Hotel. Taking the minimum 'DaysSinceLastStay' as Recency. (67 - 998)

Frequency (F): How often a customer has visited the Hotel. Taking 'DocIDHash' to count the number of reservations to get the frequency variable. (1 - 25)

Monetary (M): How much money a customer spends on the Hotel. Taking the sum of 'LodgingRevenue' and 'OtherRevenue' for monetary value. (94 - 980949.34)

Computed the RFM Scores for each customer through quartile analysis. Each quartile will give a score of 1 through 5. The final RFM segmentation will use these scores together.

Based on the generated RFM scores, we identify different segments to look at how often a customer visits us and how valuable they are. These scores and categorization will help the marketing team to create tailored services for every customer.

From RFM score (111: Core customers) to RFM score (555: Lost Customers), divided into 11 segments as follows:-

Champions, Loyals, Potential Loyalists, Promising, Need Attention, New Customers, About to Sleep, At Risk, Cannot Lose Them, Hibernating Customers, Lost Customers.

In the third phase of the project, Analysis based on the results obtained is done, as shown in the next section, and lastly concluded with meaningful insights and inferences.

EXPERIMENTAL RESULTS AND ANALYSIS

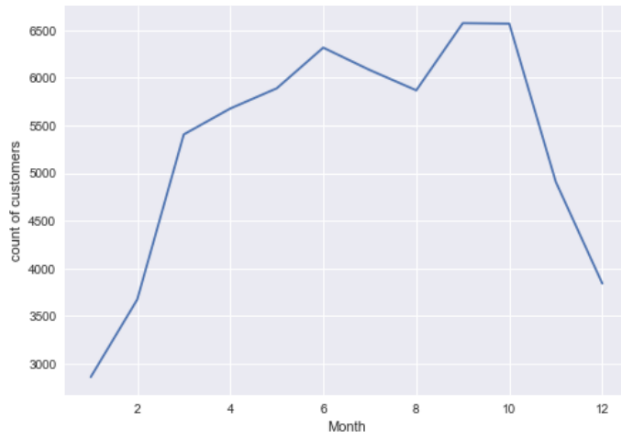


Fig.3.1: Count of Customers Monthwise

This bar graph depicts the number of people visiting the hotel per year.

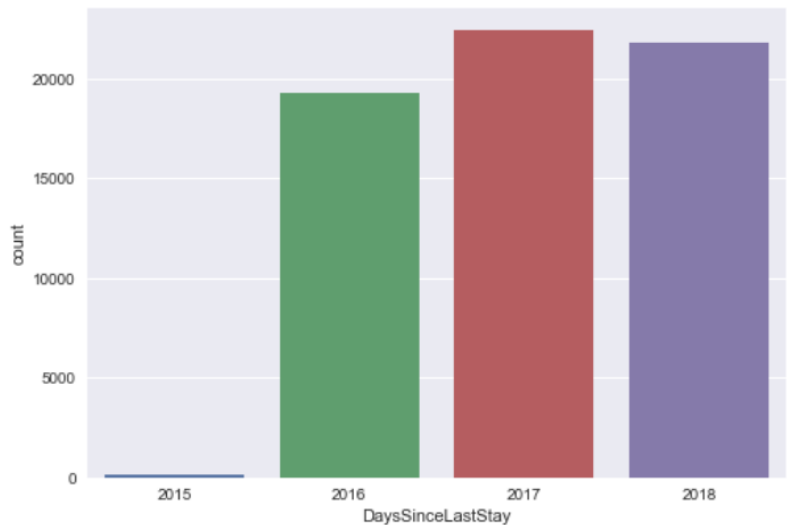


Fig.3.2: Customers visiting per year



This graph visualises the distribution of the number of rooms booked by a customer at the hotel in a single transaction.

Fig.3.3: Rooms Booking per person

Similarly this graph here visualises the percentage distribution of the number of people for which the booking has been confirmed in a single transaction

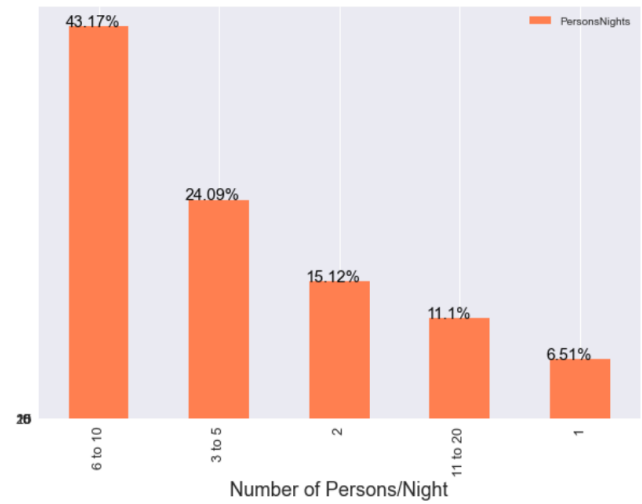


Fig.3.4: Number of Persons per Night

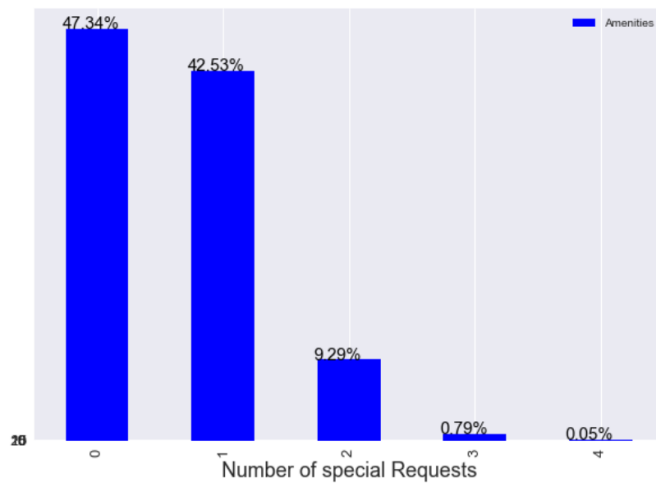


Fig.3.5: Percentage of Amenities

The bar graph here represents the diffusion of customers who require a certain special amenity when staying at the hotel.

The number of customers belonging to each cluster

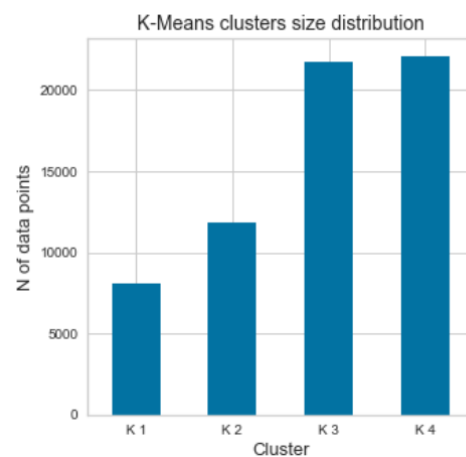
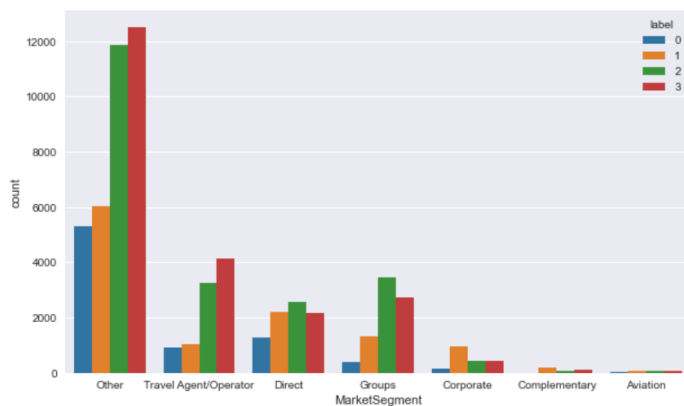
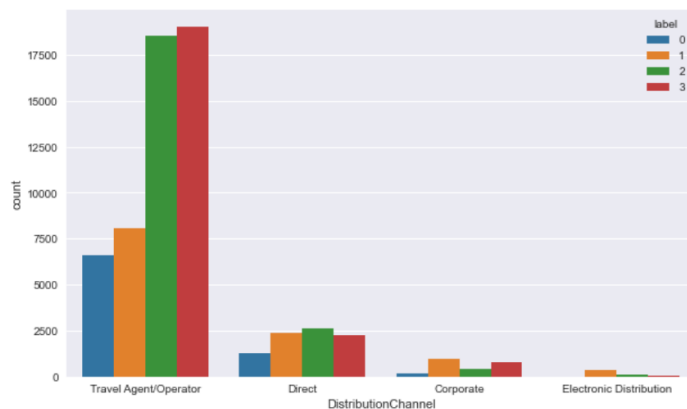


Fig.3.6: Data points per Cluster using k means

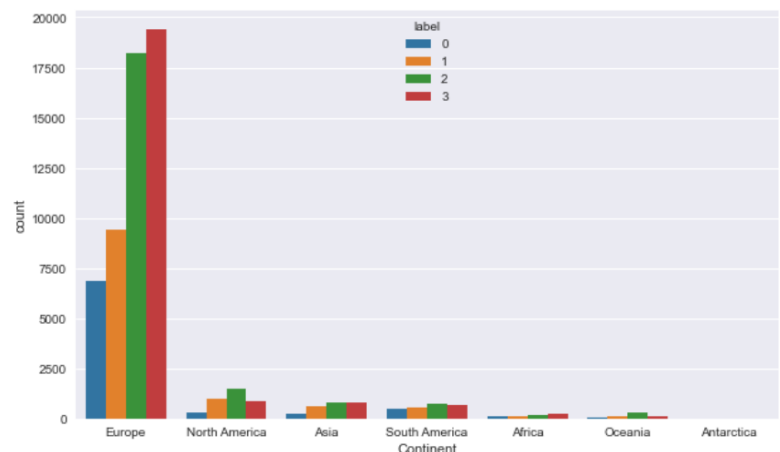


This graph portrays the number of people belonging to each cluster with respect to the market segment they belong to

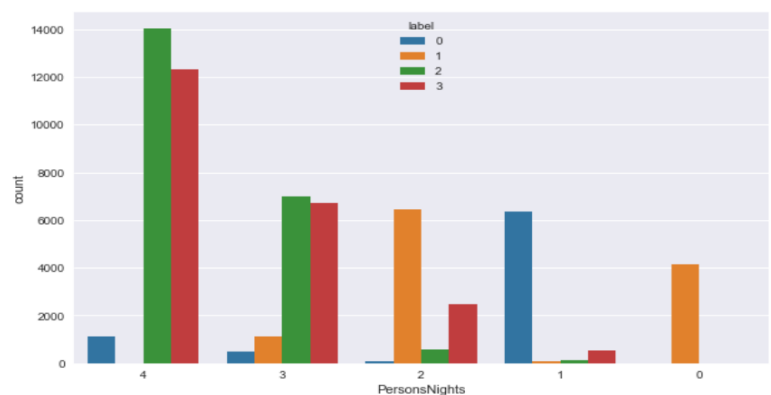


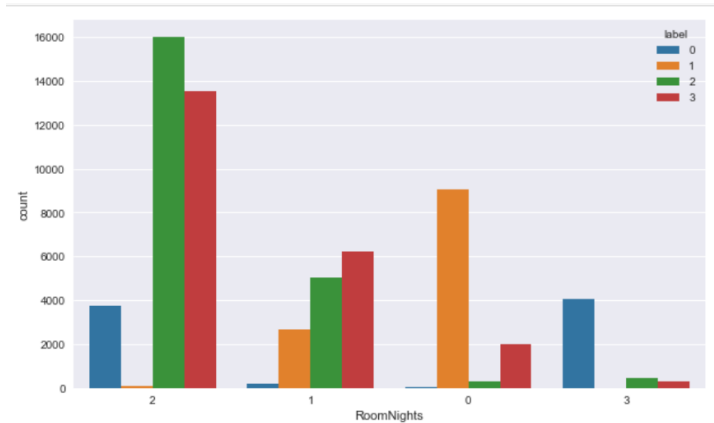
Here the graph representations the number of people present in each cluster with respect to the distribution channel through which they've got their bookings done

This graph here shows the continent wise distribution of the customers that are coming to stay at the hotel



Here the graph depicts the distribution of the customers based on the fact that booking is done for how many people. And based on that cluster wise number of people are shown





And here the graph depicts the number of people in each cluster with respect to the number of rooms that they've booked over a single transaction

This graph is showing the number of people cluster wise of each age that has visited the hotel in these three years

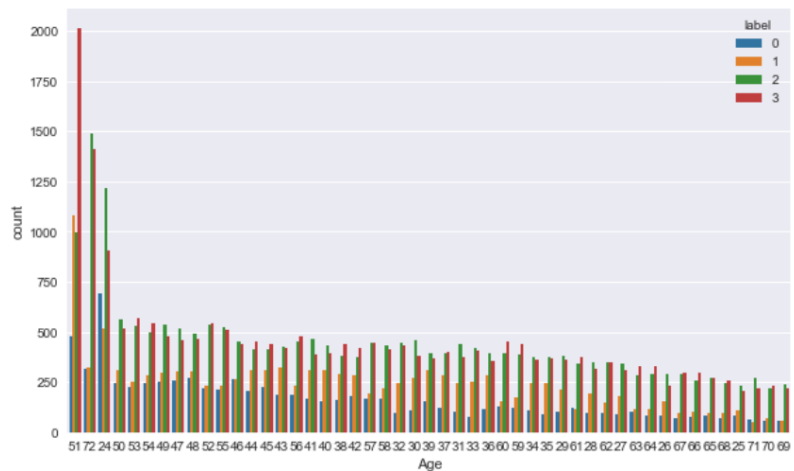


Fig.3.7: Clusters per Age

Number of customers present in each cluster with respect to the RFM segments

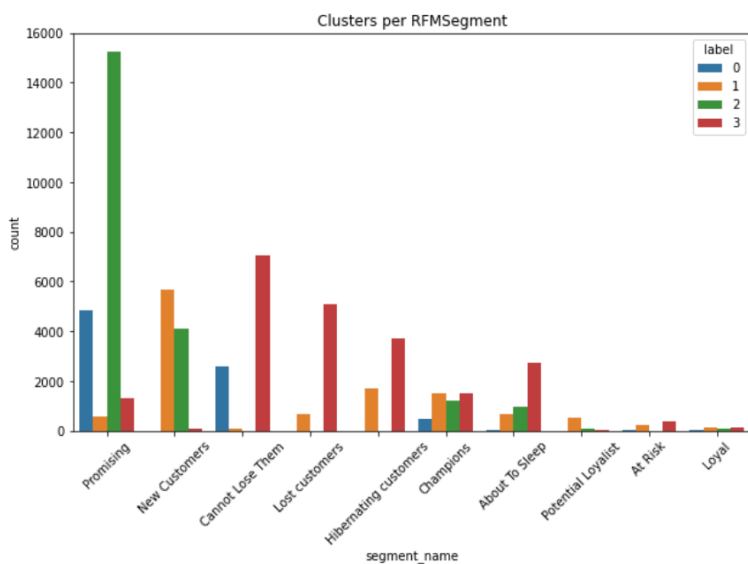


Fig.3.8: Clusters per RFM Segment

Analysis/Interpretation

1. Cluster 0: High Spenders

This cluster consists of the customers who spend lavishly. The majority of the customers in this segment spend around 851-1050 Euros as lodging revenue and 250+ Euros as other revenue. Around 40% of the customers are aged between 31-50. The customers prefer to come in a group as they book rooms for 12-14 persons' stay. More than 50% of the customers like to book rooms for 6+ days. The majority of the bookings at the hotel for customers in this segment are done through Travel Agents. The average lead time is the least for this cluster, i.e. the average number of days elapsed between the customer's booking date and arrival date is less. Customers don't tend to book their rooms well in advance.

2. Cluster 1: Low Spenders

This cluster consists of people who do not like spending money unnecessarily. More than 63% of the customers in this segment spend only 50-150 Euros on lodging revenue whereas around 80% spend less than 30 Euros on other expenses. Around 48% of the customers are aged between 31-50. The customers prefer to come solo or with their spouses as they book rooms for 1-2 persons' stay only. More than 76% of the customers like to book rooms for 1 day. The majority of the bookings at the hotel for customers in this segment are done through Travel Agents. The average lead time is the second least for this cluster. Customers don't tend to book their rooms well in advance. Some customers also tend to cancel their bookings.

3. Cluster 2: Moderate Customers

This cluster mainly consists of people who don't spend a lot of money. They spend an average amount and hence are being called moderate customers. Almost half of the people in this cluster spend 250-450 Euros as lodging revenue. 40% of people belonging to this cluster lie between the ages of 31-50. Most of the book 3 rooms for the stay of 6 people and the booking is done via Travel Agents.

4. Cluster 3: Less Frequently Visiting Customers

This cluster comprises those customers who don't revisit the hotel for a very long time and are thus very less frequent. They usually choose travel agents as their distribution channel. People aged around 50 forms a major part of this cluster. More than 50% of these customers spend money as lodging revenue in the range of 150-350 Euros and so are low spenders as well.

CONCLUSION AND FUTURE SCOPE

Organisations must develop a thorough understanding of their customers' qualities, behaviours, demographics, and other factors. Many techniques have been developed in this context. Businesses can gain complete insight into their customers using a variety of models and algorithms. This will help them quickly design relevant and unique consumer strategies by grouping customers based on their characteristic data. To demonstrate the same, a Hotel customer dataset having over three years of data was chosen. The process of dimensionality reduction using PCA was performed on the raw data and some dataset attributes were normalised. K-Means Clustering algorithm and Expectation-Maximisation (EM) Algorithm were applied to the preprocessed data. The Clustering Performance was evaluated using Internal indices such as the Silhouette index. The best results were obtained by the K-means clustering algorithm. Then RFM modelling was performed on the reduced dataset and customers were divided based on their RFM scores. Using the elbow method for finding optimal cluster numbers, four clusters were formed. The RFM scores were divided into 11 categories. Then using the DocHashID of each customer, we analysed the number of customers in each cluster belonging to different RFM categories.

In future work, the process of classification and regression can be applied to the dataset due to the high number of variables enclosed. Association rule mining can also be applied because it finds interesting associations and relationships among large sets of data items (here, customers). It allows for identifying relationships between the hotel facilities that people prefer frequently. Given a set of transactions, rules can be found that can predict the choice of an amenity based on the occurrences (choices) of other hotel services in the transaction.

REFERENCES

Research papers:

- [1] Alkhayrat et al. J Big Data (2020), “A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA” <https://doi.org/10.1186/s40537-020-0286-0>
- [2] Antonio, Nuno; de Almeida, Ana; Nunes, Luis (2020), “Lisbon, Portugal, hotel’s customer dataset with three years of personal, behavioural, demographic, and geographic information”, Mendeley Data, V1, doi: 10.17632/j83f5fsh6c.1
- [3] A.J. Christy et al. / Journal of King Saud University,” RFM ranking – An effective approach to customer segmentation”- Computer and Information Sciences 33 (2021) 1251–1257.
- [4] Abirami, M. & Pattabiraman, V. (2016). Data mining approach for intelligent customer behavior analysis for a retail store. In: proceedings of the 3rd international symposium on big data and cloud computing challenges (ISBCC–16). 283-291.
- [5] Ansari, A. & Riasi, A. (2016). Taxonomy of marketing strategies using bank customers’ clustering. International Journal of Business and Management, 11(7), 106-119.
- [6] Zhiyuan Yao, Tomas Eklund, Barbro Back (2010 IEEE),” Using SOM-Ward clustering and predictive analytics for conducting customer segmentation”,978-0-7695-4257-7/10 \$26.00 © 2010 IEEE DOI 10.1109/ICDMW.2010.121

Online:

- [1] International Standards Organization, ISO country codes 3166-3:2013, (2013). <https://www.iso.org/obp/ui/#iso:std:iso:3166:-3:ed-2:v1:en,fr> (accessed March 24, 2018).
- [2] A.J. Menezes, P.C. van Oorschot, S.A. Vanstone, Handbook of Applied Cryptography, 5th Ed., CRC Press, 1996
- [3] <http://dx.doi.org/10.17632/j83f5fsh6c.1>