Data in Brief 33 (2020) 106583

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A hotel's customers personal, behavioral, demographic, and geographic dataset from Lisbon, Portugal (2015–2018)

Nuno Antonio [a,b,*], Ana de Almeida [c,e], Luís Nunes [c,d,e]

[a] Nova IMS, Universidade Nova de Lisboa, Lisbon, Portugal
[b] CITUR, Faro, Portugal
[c] Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
[d] Instituto de Telecomunicações, Lisbon, Portugal
[e] ISTAR-Iscte, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 12 October 2020
Revised 19 November 2020
Accepted 20 November 2020
Available online 24 November 2020

Keywords:
Classification
Clustering
Data mining
Data science
Hospitality
Machine learning
Regression
RFM modeling

ABSTRACT

This data article describes a hotel customer dataset with 31 variables describing a total of 83,590 instances (customers). It comprehends three full years of customer behavioral data. In addition to personal and behavioral information, the dataset also contains demographic and geographical information. This dataset contributes to reducing the lack of real-world business data that can be used for educational and research purposes. The dataset can be used in data mining, machine learning, and other analytical field problems in the scope of data science. Due to its unit of analysis, it is a dataset especially suitable for building customer segmentation models, including clustering and RFM (Recency, Frequency, and Monetary value) models, but also be used in classification and regression problems.

* Corresponding author.
E-mail address: nantonio@novaims.unl.pt (N. Antonio).

https://doi.org/10.1016/j.dib.2020.106583

Specifications Table

Subject	Tourism, Leisure and Hospitality Management
Specific subject area	Marketing, Customer Relationship Management, and Revenue Management
Type of data	Text file and Python notebook
How data were acquired	Extraction from the hotel Property Management System (PMS) SQL database
Data format	Mixed (raw and pre-processed)
Parameters for data collection	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 guaranty 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.
Description of data collection	Data were extracted via TSQL queries executed in the production server, using Microsoft SQL Studio Manager. Python was employed to perform summary statistics.
Data source location	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.
Data accessibility	Data is available from http://dx.doi.org/10.17632/j83f5fsh6c.1

Value of the Data

- This multidimensional real-world business hotel customer dataset is of extreme utility for customer segmentation/clustering research and education.
- Students and educators are able to use the data to the study and teaching of data mining, business analytics, machine learning, and statistics. Researchers can use this dataset to develop or benchmark data mining or machine learning algorithms.
- Due to its real-world origin, this dataset contains quality issues difficult to find in classical datasets for teaching or in curated datasets. Thus, this dataset is also instrumental for students to understand and practice data cleansing and data preparation before modeling of the data.
- Due to the unit of analysis (the customer), the dataset is mostly adequate for training in the development of customer segmentation models, including clustering and RFM (Recency, Frequency, and Monetary value)
- Data is both suitable for unsupervised problems (clustering) and for supervised problems (classification and regression) due to the high number of variables enclosed.

1. Data Description

The dataset available in the supplementary file HotelCustomersDataset.tsv is provided in tab-separated value format. The unit of analysis of the dataset being the customer, each instance (row) of the dataset represents one customer. Each variable (column) represents a characteristic or description of the customer. The dataset's columns are described in Table 1.

Tables 2–4 present summary statistics of the datasets.

Fig. 1 present histograms for all numeric and Boolean variables. These statistics were obtained using Python, by using the Jupyter notebook provided as a supplementary file.

Table 1
Variables description (in the order shown in the dataset).

Variable	Type	Description
ID	Numeric	Customer ID
Nationality	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [1]
Age	Numeric	Customer's age (in years) at the last day of the extraction period.
DaysSinceCreation	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)
NameHash	Categorical	Name of the customer's SHA2–256 hash string. A hash-string is the string resulting from a mathematical function that maps a string of arbitrary length to fixed-length [2]. Hash functions are used for different purposes. In this case, to allow customer's anonymization.
DocIDHash	Categorical	SHA2–256 hash-string of the identification document number the customer provided at check-in (passport number, national ID card number, or other)
AverageLeadTime	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
LodgingRevenue	Numeric	Total amount spent on lodging expenses by the customer (in Euros). This value includes room, crib, and other related lodging expenses
OtherRevenue	Numeric	Total amount spent on other expenses by the customer (in Euros). This value includes food, beverage, spa, and other expenses
BookingsCanceled	Numeric	Number of bookings the customer made but subsequently canceled (the costumer informed the hotel he/she would not come to stay)
BookingsNoShowed	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)
BookingsCheckedIn	Numeric	Number of bookings the customer made, and which end up with a staying
PersonsNights	Numeric	The total number of persons/nights that the costumer stayed at the hotel. This value is calculated by summing all customers checked-in bookings' persons/nights. Person/nights of each booking is the result of the multiplication of the number of staying nights by the sum of adults and children
RoomNights	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
DaysSinceLastStay	Numeric	The number of days elapsed between the last day of the extraction and the customer's last arrival date (of a checked-in booking). A value of −1 indicates the customer never stayed at the hotel
DaysSinceFirstStay	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
DistributionChannel	Categorical	Distribution channel usually used by the customer to make bookings at the hotel
MarketSegment	Categorical	Current market segment of the customer
SRHighFloor	Boolean	Indication if the customer usually asks for a room on a higher floor (0: No, 1: Yes)
SRLowFloor	Boolean	Indication if the customer usually asks for a room on a lower floor (0: No, 1: Yes)
SRAccessibleRoom	Boolean	Indication if the customer usually asks for an accessible room (0: No, 1: Yes)
SRMediumFloor	Boolean	Indication if the customer usually asks for a room on a middle floor (0: No, 1: Yes)
SRBathtub	Boolean	Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes)
SRShower	Boolean	Indication if the customer usually asks for a room with a shower (0: No, 1: Yes)

(*continued on next page*)

Table 1 (*continued*)

Variable	Type	Description
SRCrib	Boolean	Indication if the customer usually asks for a crib (0: No, 1: Yes)
SRKingSizeBed	Boolean	Indication if the customer usually asks for a room with a king-size bed (0: No, 1: Yes)
SRTwinBed	Boolean	Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes)
SRNearElevator	Boolean	Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes)
SRAwayFromElevator	Boolean	Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes)
SRNoAlcoholInMiniBar	Boolean	Indication if the customer usually asks for a room with no alcohol in the mini-bar (0: No, 1: Yes)
SRQuietRoom	Boolean	Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes)

Table 2
Summary statistics - Categorical variables.

Variable	Count	Unique	Max. frequency
Nationality	83,590	188	12,422
NameHash	83,590	80,642	47
DocIDHash	83,590	76,993	3657
DistributionChannel	83,590	4	68,569
MarketSegment	83,590	7	48,039

Table 3
Summary statistics - Boolean variables.

Variable	Count	No (0)%	Yes (1)%
SRHighFloor	83,590	95.53	4.74
SRLowFloor	83,590	99.86	0.14
SRAccessibleRoom	83,590	99.97	0.03
SRMediumFloor	83,590	99.91	0.09
SRBathtub	83,590	99.97	0.03
SRShower	83,590	99.82	0.18
SRCrib	83,590	98.68	1.32
SRKingSizeBed	83,590	64.73	35.27
SRTwinBed	83,590	85.74	14.26
SRNearElevator	83,590	99.97	0.03
SRAwayFromElevator	83,590	99.64	0.36
SRNoAlcoholInMiniBar	83,590	99.99	0.01
SRQuietRoom	83,590	91.16	8.84

Table 4
Summary statistics - Numeric variables.

Variable	Count	Mean	Std	Min.	25%	50%	75%	Max.
ID	83,590	41,795.5	24,130.5	1	20,898.2	41,795.5	62,692.8	83,590
Age	79,811	45.398	16.5724	−11	34	46	57	122
DaysSinceCreation	83,590	453.641	313.39	0	177	397	723	1095
AverageLeadtime	83,590	66.196	87.759	−1	0	29	103	588
LodgingRevenue	83,590	298.802	372.852	0	59	234	402	21,781
OtherRevenue	83,590	67.5891	114.328	0	2	38.5	87.675	7730.25
BookingsCanceled	83,590	0.00202	0.06677	0	0	0	0	9
BookingsNoShowed	83,590	0.00063	0.02955	0	0	0	0	3
BookingsCheckedIn	83,590	0.79462	0.69578	0	1	1	1	66
PersonsNights	83,590	4.64913	4.56767	0	1	4	6	116
RoomNights	83,590	2.35854	2.28175	0	1	2	4	185
DaysSinceLastStay	83,590	401.067	347.205	−1	26	366	693	1104
DaysSinceFirstStay	83,590	403.349	347.971	−1	27	369	697	1186

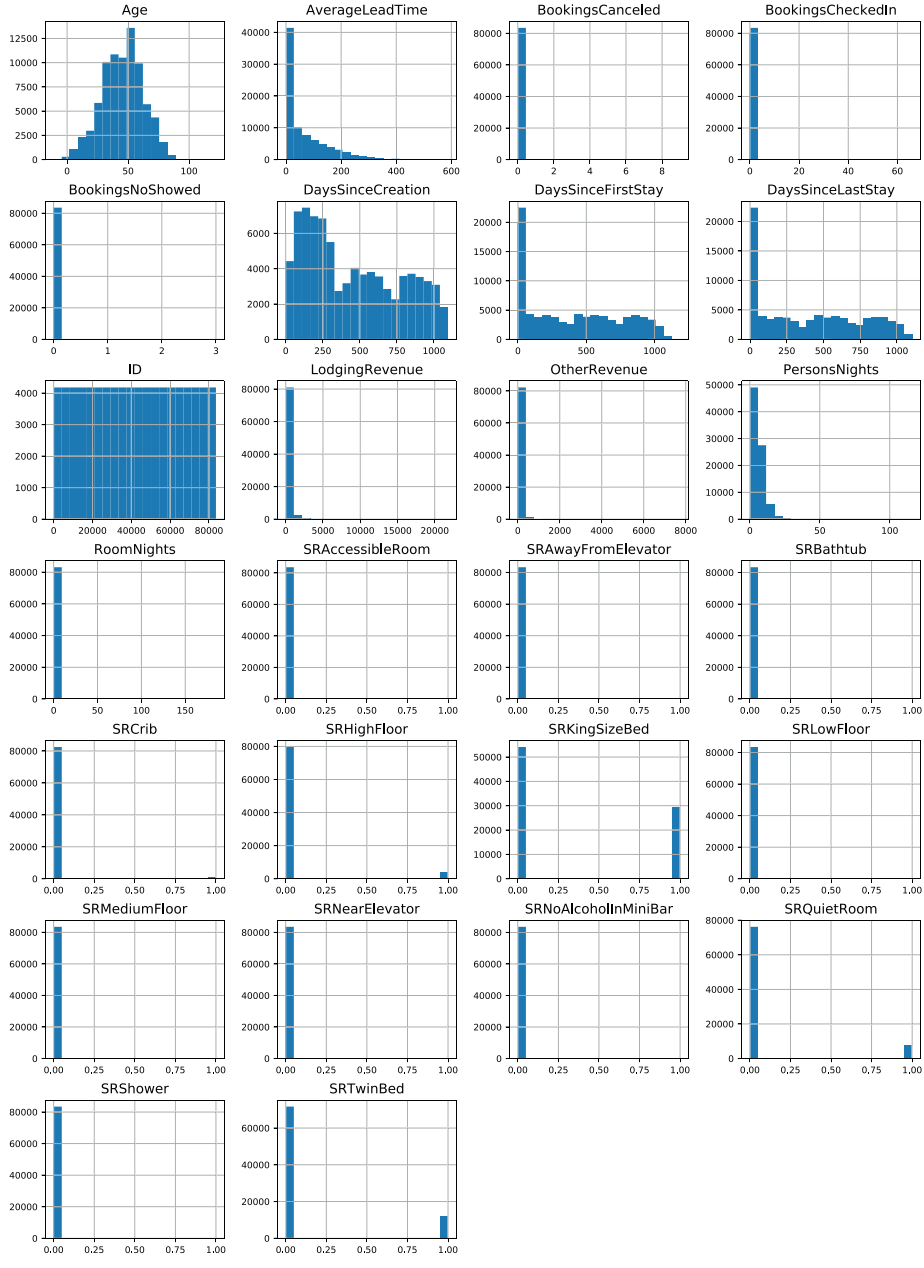

Fig. 1. Histograms of all numeric and Boolean variables.

2. Experimental Design, Materials and Methods

The main objective for the creation of this dataset was to share real-world business data that could be used in data mining, business analytics, machine learning, statistics, education and research.

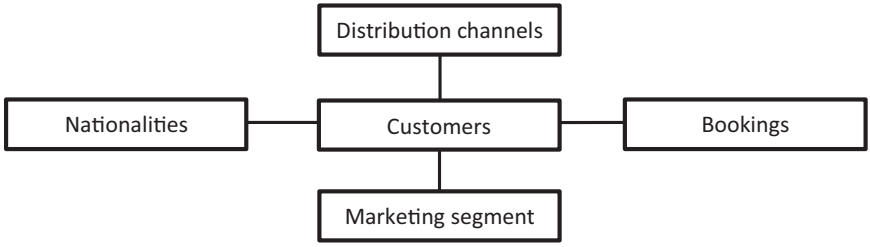

Fig. 2. Diagram of PMS database tables where variables were extracted from.

Data were extracted from the hotel PMS database server by executing a TSQL query on SQL Server Studio Manager, the integrated environment tool for managing Microsoft SQL databases [3]. Data were aggregated to the unit of analysis, the customer, so that it could be used to work in a variety of marketing and customer relationship management problems, such as customer segmentation, customer lifetime value calculation, campaign planning, or pattern analysis. As illustrated in Fig. 2, the aggregation was made with data from different database tables, namely:

- Customers – the base table, containing personal data, including geographic and demographic characteristics of the customer.
- Nationality – table with the description of the country of the nationality of the customer.
- Distribution channel – table with the description of the distribution channel.
- Market segment – table with the description of the market segment.
- Bookings – a table that contains behavioral characteristics of the customer, including its reservation and spending details.

Since hospitality operations are characterized by seasonality, it was decided to extract three full years of customer behavioral data: from 2015 to 2018. All time-based variables, such as when the customer was created in the database (*DaysSinceCreation*) or the number of days since the customer last stayed at the hotel (*DaysSinceLastStay*), are a construction based on the last day of extraction - December 31, 2018.

Domain knowledge is essential in any analytical field, especially in data mining, business analytics, and machine learning [4,5]. Therefore, a few considerations are presented to help use the dataset, particularly in the critical data understanding and data preparation phases [6].

1. Hotels do not create a customer record for every guest. Usually, hotels only create a profile for the booking holder. However, some hotels have a policy of creating a profile for each guest companion (adult or children) only in particular cases. In Europe, a hotel can only create a profile if the customer gives express authorization.
2. Typically, a customer profile is created in one of three moments: at the customer's first checked-out at the hotel, at the customer first cancelation, or the customer's first no-show.
3. Due to PMS application malfunctions, user errors, customers providing different identification documents or use other names (e.g., first plus last name or full name), sometimes hotels have more than one profile for the same customer.
4. Usually, only one personal information is necessary to make a booking at a hotel: the booking holder's name.
5. Only after a customer's first stay can hotels confirm the guest's personal details, such as nationality, identification card number, and birthdate, among others.

Ethics Statement

Given that the hotel providing the data requested anonymity and due to the preservation of personal data, such as name, ID number and date of birth, every effort was made to anonymize

or transform fields that directly or indirectly allowed identification of the hotel or the hotel's customers. Nevertheless, it was ensured that the data maintained its original properties, such as the existence of duplicates, outliers, missing values, among others. In this way, the authors guarantee the anonymity and privacy of the data, without influencing its quality for the appropriate types of use.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

The authors would like to thank the hotel for allowing their data to be shared publicly.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2020.106583.

References

[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] Microsoft, SQL Server Management Studio (SSMS), (2017). https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms (accessed March 24, 2018).
[4] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd Ed., Elsevier, Waltham, MA, USA, 2012 http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf (accessed May 7, 2019).
[5] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (2012) 78–87.
[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: step-by-step data mining guide, Model. Agency (2000) https://the-modeling-agency.com/crisp-dm.pdf. (accessed September 10, 2015).