Text

Description automatically generated

DATA SCIENCE AND MACHINE LEARNING

**BookMe Project**

April 17th, 2022

**Data Science and Machine Learning**

**BookMe Project**

***Group Members:***

Mariana Dias Gil, *nº 20211043*

Pedro Bernardes, *nº20211774*

Ricardo Daniel Pacheco De Almeida, *nº 20201335*

Soraia Andreia Anselmo Alves, *nº 20201840*

Table of Contents

[Table of Methodologies’ Description 4](#_Toc100600316)

[Table of Figures 5](#_Toc100600317)

[1. Introduction 6](#_Toc100600318)

[2. Problem Definition 6](#_Toc100600319)

[3. Data Exploration and Preparation 6](#_Toc100600320)

[3.1. Data Exploration 6](#_Toc100600321)

[2.1. Data Preparation 7](#_Toc100600322)

[3. Data Processing 8](#_Toc100600323)

[3.1. Clustering 8](#_Toc100600324)

[3.1.1. Customer Characteristics 8](#_Toc100600325)

[3.1.2. Customer Satisfaction Level 8](#_Toc100600326)

[3.1.3. Customer Attrition 8](#_Toc100600327)

[3.2. Results 8](#_Toc100600328)

[3.3. Discussion 8](#_Toc100600329)

[4. Conclusion 8](#_Toc100600330)

[5. Bibliography 9](#_Toc100600331)

[6. Annexes 10](#_Toc100600332)

[6.1. Figures 10](#_Toc100600333)

[2.1. Methodologies 17](#_Toc100600334)

# Table of Methodologies’ Description

[Methodology 1 – KNN 17](#_Toc100600227)

[Methodology 2 - PCA Analysis 17](#_Toc100600228)

[Methodology 3 – K-Means Clustering 18](#_Toc100600229)

[Methodology 4 – Hierarchical clustering 18](#_Toc100600230)

# Table of Figures

[Figure 1 - Summary of dataset's attributes and description 11](#_Toc100944649)

[Figure 2 - First three entries of BookMe dataset 11](#_Toc100944650)

[Figure 3 - Duplicated entries removed 12](#_Toc100944651)

[Figure 4 - Information on BookMe dataset 12](#_Toc100944652)

[Figure 5 - Descriptive statistics for numerical variables 13](#_Toc100944653)

[Figure 6 - Countplot of "RoomType" and "TypeTravel" 13](#_Toc100944654)

[Figure 7 - Countplot of "Longevity" and "Churn" 13](#_Toc100944655)

[Figure 8 - Skweness values 14](#_Toc100944656)

[Figure 9 - Kurtosis values 14](#_Toc100944657)

[Figure 10 - Boxplots for 'Amenities' and 'Staff' 14](#_Toc100944658)

[Figure 11 - Boxplots for 'PriceQuality' and 'CheckOut' 14](#_Toc100944659)

[Figure 12 - Boxplot for 'Cleanliness' 15](#_Toc100944660)

[Figure 13 - Missing values distribution (after transformation of underage people to missing) 15](#_Toc100944661)

[Figure 14 - 'Year\_Birth' descriptive statistics before imputation 15](#_Toc100944662)

[Figure 15 - 'Year\_Birth' descriptive statistics after imputation 15](#_Toc100944663)

[Figure 16 - 'Wifi' values before correction 15](#_Toc100944664)

[Figure 17 - 'Wifi' values after correction 16](#_Toc100944665)

[Figure 18 - Distribution of new variable 'Gender' 16](#_Toc100944666)

[Figure 19 - Distribution of new variable 'Age' 16](#_Toc100944667)

[Figure 20 - Distribution of new variable 'satisfaction\_avg' 16](#_Toc100944668)

[Figure 21 - Heatmap 17](#_Toc100944669)

[Figure 22 - PCA Analysis 17](#_Toc100944670)

[Figure 23 – Customers’ Clusters 18](#_Toc100944671)

[Figure 24 - Customers’ Clusters: Age Distribution 18](#_Toc100944672)

[Figure 25 - Customers’ Clusters: Descriptive Statistics 19](#_Toc100944673)

[Figure 26 - Customers' Satisfaction Clusters 19](#_Toc100944674)

[Figure 27 - Customers' Satisfaction Clusters: Accommodation 20](#_Toc100944675)

[Figure 28 - Customers' Satisfaction Clusters: Reception 20](#_Toc100944676)

[Figure 29 - Accommodation vs Reception 20](#_Toc100944677)

[Figure 30 - Reception vs Catering 20](#_Toc100944678)

[Figure 31 - Customers' Satisfaction Clusters: Descriptive Statistics 21](#_Toc100944679)

[Figure 32 - Clients' Retention Level 21](#_Toc100944680)

[Figure 33 - Average Satisfaction 21](#_Toc100944681)

[Figure 34 - Clients' Retention Level: Descriptive Statistics 22](#_Toc100944682)

# Introduction

The BookMe company is a well-established company operating in the hospitality sector, that provides accommodation to tourists and travellers, delivering necessary lodging services to those who travel the world, whether for leisure or business.

The accommodations can only be booked online, through their website. Currently, the company has around 30.000 registered clients and serves more than 100.000 customers a year. The main goal of BookMe company is to provide the best conditions as possible, so by the end of the stay, each customer receives a survey to evaluate their satisfaction level (within a rate of 0 to 5), over the different services provided: location, price, amenities, among others.

Although the company presented stable results for the last three years, it is concerned about its financial projections for the next triennium. This report aims to address the volatility’s profit growth perspectives for the next three years by customer retention,.

The Data Science team doesn’t have a simple goal for the analysis and are not interested in prediction (*eg* budgeting prediction), as the aim is more related to discover reviling things about the company’s customers features (unsupervised learning). Is there an enlighten way to visualize the data? Can the data be divided in subgroups among the features or observations?

Summarize the basic results and conclusions that you will present

# Problem Definition

The main objective of the project is to find a structure in the collection of the company’s customer characteristics, *i.e.*, clustering to perform market segmentation. Clustering refers to a very broad set of techniques for finding subgroups whose members are similar in some way (or clusters), in a dataset.

There are a great number of clustering methods. The team is going to focus on the following approaches: (i) K-Means Clustering and (ii) Hierarchical Clustering Each one of them has its advantages and disadvantages, which will be identified throughout the report.

# Data Exploration and Preparation

BookMe provided a customer’s database with features about the costumers’ characteristics and business features, as identified in Figure 1. While the company explained that each customer receives a survey, it’s clear that the company needs to improve its reinforcement, as the database provided only has 15,5 thousand customers contrasting with the 100 thousand a year and may not be truly representative.

For the purpose of this project, the team has used Jupyter notebook and python libraries like pandas, numpy, matplotlib, seaborn and sklearn.

As first steps, the team imported the file *'data\_cluster.csv'* to the Jupyter notebook (please refer to Figure 2) and proceeded to remove the duplicate entries, as visible in Figure 3. At this moment, final dataset was composed by a total of 15,586 rows/observations and 21 columns/attributes – 16 numerical attributes and 5 categorical (Figure 4).

## Data Exploration

The analysis started with the exploration of the descriptive statistics of the numerical variables of the dataset (Figure 5). The team was able to conclude the following:

1. **Satisfaction Level:** the average satisfaction level with the different services provided by BookMe company is ‘3’. Apart from ‘Location’, ‘Staff’, ‘PriceQuality’, ‘CheckOut’, ‘Checkin’ and ‘Cleanliness’, all the other services have been rated with ‘0’, at least once. On a more positive note, all of them have also been rated with ‘5’, at least once;
2. **Customer Segmentation:** around 80% of the are registered in the platform for about 1 year, at least, and more than 50% of them are in a ‘nchurn’ situation, which gives a positive rate. Almost 70% of the clients uses this hospitality service for business purposes and the most booked ‘RoomType’ is ‘single’ with close to 40% preference. When combined with the ‘TravelType’, the team found a pattern in which the ‘single’ and ‘suite’ room types were mostly booked by travellers in business and the ‘double’ ones were the preference for leisure travels (Figure 6);
3. **Retention (Churn Level):** the most recent clients, *ie*, the ones registered in the BookMe platform for less than 1 year, presented more churn (Figure 7). The team considers that this finding is worth more exploration, because it could mean that the platform user experience could be improved, or it is necessary to create new incentives (*eg* discounts) for newcomers.

Regarding the shape and density of the distributions underlying the variables, the “RewardPoints” variable had the highest standard deviation, which indicated that these values were spread out over a wider range. However, apart from some moderate skewness (Figure 8) values, related with the variables “Amenities”, “Staff”, “PriceQuality”, “CheckOut” and “Cleanliness”, as per the kurtosis (Figure 9) calculations, the team did not identify the presence of any potential outliers, even though it explored this possibility, using some visual techniques as per Figure 10, Figure 11 and Figure 12. The variable ‘PriceQuality’ presented some values highly deviated from the normal distribution, however they corresponded to 1,610 clients, which means more than 10% of the data. As per the rule of thumb, no more than 3% of the outliers should be removed, so all the registers were kept, since there was no way of differentiating them.

In terms of issues within the customers database, the team identified 195 missing values in the customers’ year of birth, as well as 1.051 clients with less than 18 years old, being that the youngest register belongs to a person born in 2004, meaning with an age of 8. As per the business rules, the survey is only provided to registered users (assuming that only adults can register), after each stay. Additionally, the maximum rate for ‘Wifi’ was appointed as ‘6’, which is clearly an incoherence (the scale is 0-5). The last incoherence identified was related with the longevity of stay, which had three responses (‘yes’, ‘y’, ‘no’), instead of two (this problem was instantly resolved by assuming ‘y’ as ‘yes’).

## Data Preparation

To outline the issues and incoherencies found earlier, the team proceeded to the preprocess of the data. Concerning the underage customers, the team replaced the values with missing, adding to the original missing values, leaving 1.246 entries to fill in (Figure 13). As the database is not very extensive, the team preferred to treat the missing values, instead of dropping them. Even though the variable did not present a high correlation with any other, the team applied the KNN (Methodology 1) Imputer technique to fill in the missing values. After comparing the descriptive statistics of the variable ‘Year\_Birth’ distribution, before and after the imputation, it was concluded that the result was good, because the main values have pretty much remained equal, and the standard deviation decreased a little (Figure 14 and Figure 15). Regarding misclassifications identified in the variable ‘Wifi’, the team replaced the 36 entries with a rate of ‘6’, by the maximum possible value of ‘5’ (Figure 16 and Figure 17).

Based on the available attributes it was possible to create new features, namely the gender (Figure 18), based on the prefixes ‘Mr’ and ‘Ms’ on the variable ‘Name’; age (Figure 19), calculated from difference of the current year and the year of birth (Figure 20); and average overall satisfaction (average satisfaction level from the different services).

Finally, based on a heatmap figure (Figure 21), the team was able to identify which features were more correlated with one another. The result was as follows: “Comfort” with “FoodDrink” (0.7); “ReceptionSchedule” with “Location” (0.6); “Wifi” with “OnlineBooking” and “BarService” (0.6); “Staff” with “BarService” (0.7) and “OnlineBooking” (0.6); “OnlineBooking” with “BarService” (0.7); “PriceQuality” with “Cleanliness” (0.6) and “CheckOut” with “Cleanliness” (0.6). Even though clustering doesn't rely on linear assumptions, and so collinearity wouldn't cause issues, the clustering algorithms do not present great values considering many features. As a result, the team proceeded to leverage the correlated variables via PCA (Methodology 2) for effects of dimensionality reduction – it was defined a 70% threshold for the data variance explained or until the number of PCAs was reasonable - for the choosing of the component number. As Figure 22 illustrates, the outcome did have any business meaning, so the team proceeded to unite the variables based on their correlation and meaning (type of service). Namely, the team created the following variables, which combine the average satisfaction level from their respective services: (i) Accommodation (comfort, amenities, room space. cleanliness and Wi-Fi); (ii) Reception (reception schedule, staff, online booking, checkout and check-in); (iii) Catering (food and drink and bar service). The variables corresponding to the Price Quality and Location were included without any transformation.

# Data Processing

## Clustering

Once the preprocess is done, and the set is manageable, the team proceeded to clustering and finding patterns by grouping similar instances of data together. The unsupervised learning algorithms applied were the K-Means (Methodology 3) and the Hierarchical Clustering (Methodology 4). The Hierarchical Clustering, with SSR (Methodology 5) was considered to determine the best number of clusters. After the number was computed, the team applied K-Means.

### Customer Characteristics (Figures 23 - 25): The team identified three distinct groups of customers. The variables considered were ‘Age’, ‘TypeTravel’ and ‘RoomType’.

### (i) The first group contains 46% of all clients with an average age of 43 years old and the gender is equally distributed. These clients travel in business, book single rooms and have lowest number of reward points; (ii) The second group contains 31% of all clients with an average age of 44 years old and the clients are mostly female. These clients travel for leisure, book double rooms and have the highest number of reward points; (iii) The last group contains 23% of all clients with an average age of 38 years old and is mostly female. These clients travel in business and book double rooms.

### Customer Satisfaction Level (Figures 26-31): The team identified three distinct groups of customers regarding their satisfaction level. The variables considered were ‘Accommodation’, ‘Reception’, ‘Catering’, ‘Location’ and ‘PriceQuality’.

**(i)**The first cluster contains 31% of clients, mostly male, and has the lowest satisfaction in general, in particular concerning the accommodation, reception and price quality services; **(ii)** The second cluster contains 36% of clients, mostly female, and have the highest number of reward points. This group gives the highest ratings to all services, except for price quality; **(iii)** The last group contains 34% of clients, mostly female, and have the lowest number of reward points. These clients give the lowest rates do catering and location, however, give the highest rates to price quality.

### Customer Retention (Figures 32-35): The team identified three distinct groups of customers regarding their retention level. The variables considered were ‘Churn’, ‘Longevity’ and ‘satisfaction\_avg’. Even though the variable ‘satisfaction\_avg’ is calculated from the variables used in the customer satisfaction level clusters, the team found some good results and therefore has kept this perspective.

**(i)**The first group contains 18% of clients, largely female, and presents low recency. The majority of this group have no churn, nonetheless, the team detected a few churn cases that require some attention; **(ii)** The second group contains 50% of clients, mostly female, and presents high number of reward points, as for recency. This group has no churn cases. It is worth mentioning that the satisfaction level is high for this group; **(iii)** The last group contains 32% of clients, mostly male, and presents low reward points. This group is all churn, with high recency. Additionally, the satisfaction level is the lowest.

## Segmentation

The team proceeded to the concatenation of all perspectives. To do this, for each customer was applied an ID code corresponding to the union of the clusters, *eg,* “101” (Cluster 1 in customer characteristics + Cluster 0 in customer satisfaction + Cluster 1 in customer retention).

FALAR DA TEORIA DO RICARDO

## Marketing Plan

# Conclusion

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?

# Bibliography

* James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). *An introduction to statistical learning: With applications in R*. Springer;
* Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data*. O'Reilly;
* Harrison, M. (2019). *Machine learning pocket reference: Working with structured data in Python*. O'Reilly.

# Annexes

## Figures

Figure 1 - Summary of dataset's attributes and description



Figure 2 - First three entries of BookMe dataset

Uma imagem com mesa

Descrição gerada automaticamente

Figure - Duplicated entries removed

Uma imagem com mesa

Descrição gerada automaticamente

Figure 4 - Information on BookMe dataset

Uma imagem com mesa

Descrição gerada automaticamente

Figure 5 - Descriptive statistics for numerical variables

Uma imagem com mesa

Descrição gerada automaticamente

Figure 6 - Countplot of "RoomType" and "TypeTravel"

Chart, bar chart

Description automatically generated

Figure 7 - Countplot of "Longevity" and "Churn"

Chart, bar chart

Description automatically generated

Figure - Skweness values

Uma imagem com texto

Descrição gerada automaticamente

Figure - Kurtosis values

Uma imagem com texto, recibo

Descrição gerada automaticamente

Figure 10 - Boxplots for 'Amenities' and 'Staff'

Chart, box and whisker chart

Description automatically generated

Figure 11 - Boxplots for 'PriceQuality' and 'CheckOut'

Chart, box and whisker chart

Description automatically generated

Figure - Boxplot for 'Cleanliness'

Chart

Description automatically generated

Figure 13 - Missing values distribution (after transformation of underage people to missing)

Text

Description automatically generated with medium confidence

Figure 14 - 'Year\_Birth' descriptive statistics before imputation

Text

Description automatically generated

Figure 15 - 'Year\_Birth' descriptive statistics after imputation

Text

Description automatically generated

Figure 16 - 'Wifi' values before correction

Table

Description automatically generated

Figure 17 - 'Wifi' values after correction

Table

Description automatically generated

Figure 18 - Distribution of new variable 'Gender'

Text

Description automatically generated with medium confidence

Figure 19 - Distribution of new variable 'Age'

Text, letter

Description automatically generated

Figure 20 - Distribution of new variable 'satisfaction\_avg'

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceFigure 21 - Heatmap

Figure 22 - PCA Analysis

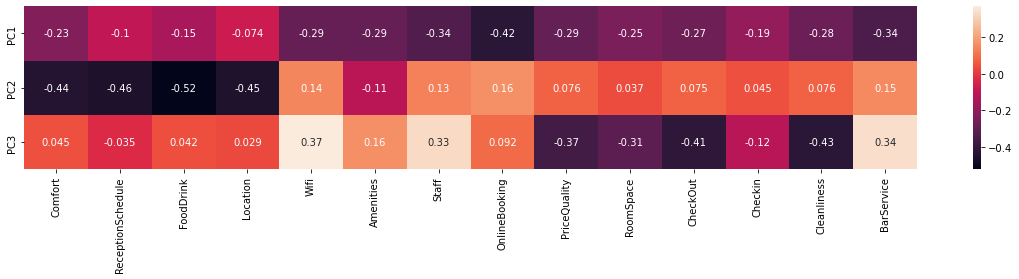


Figure 23 – Customers’ Clusters

Table

Description automatically generated

Figure 24 - Customers’ Clusters: Age Distribution

Chart, histogram

Description automatically generated

Figure 25 - Customers’ Clusters: Descriptive Statistics

Text

Description automatically generated Graphical user interface, text, application, email

Description automatically generatedText

Description automatically generated with medium confidence

Figure 26 - Customers' Satisfaction Clusters

Graphical user interface

Description automatically generated with medium confidence

Figure 27 - Customers' Satisfaction Clusters: Accommodation

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedFigure 28 - Customers' Satisfaction Clusters: Reception

Figure 29 - Accommodation vs Reception

Chart, scatter chart

Description automatically generated

Figure 30 - Reception vs Catering

Scatter chart

Description automatically generated

Figure 31 - Customers' Satisfaction Clusters: Descriptive Statistics

Table

Description automatically generatedGraphical user interface, text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Figure 32 - Clients' Retention Level

Table

Description automatically generated

Figure 33 - Average Satisfaction

Chart, histogram

Description automatically generated

Figure 34 - Clients' Retention Level: Descriptive Statistics

Graphical user interface, text

Description automatically generated Text, table

Description automatically generated Graphical user interface, text, application

Description automatically generated

## Methodologies

Methodology 1 – KNN

**K-Nearest** **Neighbours[[1]](#footnote-1)**

The KNN (k-nearest neighbours) is the most common neighbourhood-based method is k-nearest neighbours (KNN). To label each new point, KNN looks at a k number (where k is an integer value) of nearest labelled points and has these already labelled neighbours vote on how to label the new point. In Python, by default, KNN uses Euclidean distance to measure what is closest.

The choice of k is very important. If k is set to a very low value, KNN becomes very flexible, drawing highly nuanced boundaries and potentially overfitting the data. If k is set to a very high value, KNN becomes inflexible, drawing a too rigid boundary and potentially underfitting the data.

*Strengths:* Unlike linear methods, KNN is highly flexible and adept at learning more complex, nonlinear relationships. Yet, KNN remains simple and interpretable.

*Weaknesses:* KNN does poorly when the number of observations and features grow. KNN becomes computationally inefficient in this highly populated, high-dimensional space since it needs to calculate distances from the new point to many nearby labelled points in order to predict labels. It cannot rely on an efficient model with a reduced number of parameters to make the necessary prediction. Also, KNN is very sensitive to the choice of k. When k is set too low, KNN can overfit, and when k is set too high, KNN can underfit.

Methodology 2 - PCA Analysis

**Principal component analysis[[2]](#footnote-2)**

Principal component analysis (PCA) refers to the process by which principal component are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach, since it involves only a set of features X1, X2,...,Xp, and no associated response Y . Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization (visualization of the observations or visualization of the variables).

Suppose that we wish to visualize n observations with measurements on a set of p features, X1, X2,...,Xp, as part of an exploratory data analysis. We could do this by examining two-dimensional scatterplots of the data, each of which contains the n observations’ measurements on two of the features. However, there are such scatterplots. If *p* is large, then it will certainly not be possible to look at all of them; moreover, most likely none of them will be informative since they each contain just a small fraction of the total information present in the data set. Clearly, a better method is required to visualize the n observations when p is large. In particular, we would like to find a low-dimensional representation of the data that captures as much of the information as possible. For instance, if we can obtain a two-dimensional representation of the data that captures most of the information, then we can plot the observations in this low-dimensional space. PCA provides a tool to do just this. It finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the *p* features.

Methodology 3 – K-Means Clustering

**K-Means1**

*K*-means clustering is a simple and elegant approach for partitioning a data set into *K* distinct, non-overlapping clusters. To perform *K*-means clustering, we must first specify the desired number of clusters *K*; then the *K*-means algorithm will assign each observation to exactly one of the K clusters.

It optimizes the grouping by minimizing the within-cluster variation (also known as inertia) such that the sum of the within-cluster variations across all k clusters is as small as possible. To speed up this clustering process, k-means randomly assigns each observation to one of the k clusters and then begins to reassign these observations to minimize the Euclidean (or others) distance between each observation and its cluster’s center point, or centroid. As a result, different runs of k-means—each with a randomized start—will result in slightly different clustering assignments of the observations. From these different runs, we can choose the one that has the best separation, defined as the lowest total sum of within-cluster variations across all k clusters.

Methodology 4 – Hierarchical clustering

**Hierarchical Clustering1**

An alternative clustering approach—one that does not require us to precommit to a particular number of clusters—is known as hierarchical clustering. One version of hierarchical clustering called agglomerative clustering uses a tree-based clustering method and builds what is called a dendrogram. A dendrogram can be depicted graphically as an upside-down tree, where the leaves are at the bottom and the tree trunk is at the top.

The leaves at the very bottom are individual instances in the dataset. Hierarchical clustering then joins the leaves together—as we move vertically up the upside-down tree—based on how similar they are to each other. The instances (or groups of instances) that are most similar to each other are joined sooner, while the instances that are not as similar are joined later. With this iterative process, all the instances are eventually linked together forming the single trunk of the tree.

This vertical depiction is very helpful. Once the hierarchical clustering algorithm has finished running, we can view the dendrogram and determine where we want to cut the tree—the lower we cut, the more individual branches we are left with (i.e., more clusters). If we want fewer clusters, we can cut higher on the dendrogram, closer to the single trunk at the very top of this upside-down tree. The placement of this horizontal cut is similar to choosing the number of k clusters in the k-means clustering algorithm.

1. Patel, A. A. (2019). *Hands-on unsupervised learning using Python: How to build applied machine learning solutions from unlabeled data*. O'Reilly [↑](#footnote-ref-1)
2. James, G., Witten, D., Hastie, T. J., & Tibshirani, R. J. (2017). An introduction to statistical learning: With applications in R. Springer [↑](#footnote-ref-2)