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*

Ricardo Daniel Pacheco De Almeida, *nº 20201335*

Soraia Andreia Anselmo Alves, *nº 20201840*

Table of Contents

[1. Introduction 4](#_Toc100514908)

[2. Problem Definition 4](#_Toc100514909)

[3. Data Exploration and Preparation 4](#_Toc100514910)

[3.1. Data Exploration 4](#_Toc100514911)

[2.1. Data Preparation 5](#_Toc100514912)

[3. Experimental Evaluation 6](#_Toc100514913)

[3.1. Algorithms 6](#_Toc100514914)

[3.2. Results 6](#_Toc100514915)

[3.3. Discussion 6](#_Toc100514916)

[4. Conclusion 6](#_Toc100514917)

[5. Bibliography 7](#_Toc100514918)

[6. Annexes 8](#_Toc100514919)

[6.1. Figures 8](#_Toc100514920)

[2.1. Methodologies 15](#_Toc100514921)

# 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

optimizing the marketing department annual budget, this means*,* creating more efficient campaigns. The approach taken by the Data Science team is the identification of segments within the company’s Customer database with different perspectives: customer satisfaction, customer characteristics, xxxxx to target the marketing activities.

The Data Science team doesn’t have a simple goal for the analysis and are not interested in prediction (*eg* budgeting prediction), as they 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, (ii) Hierarchical Clustering, (iii) xxxx. 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. **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 skweness (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, the team still have 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 overage people 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). To delineate this question, the team compound the 14 features evaluated and calculate the corresponding average satisfaction, per client. 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 and defined a 60% threshold for the data variance explained for the choosing of the component number. As Figure 22 illustrates,

# Experimental Evaluation

## Algorithms

## Results

## Discussion

# 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;

# 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'

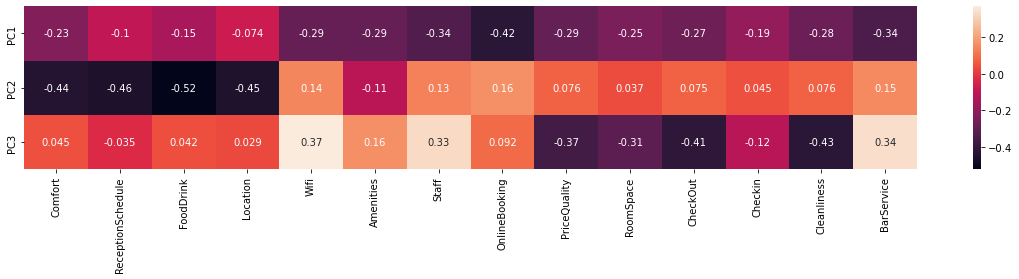
Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceFigure - Heatmap

Figure - PCA Analysis



## Methodologies

Methodology 1 – KNN

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

The KNN (k-nearest neighbors) is the most common neighborhood-based method is k-nearest neighbors (KNN). To label each new point, KNN looks at a k number (where k is an integer value) of nearest labeled points and has these already labeled neighbors 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 labeled 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.

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)