A report that describes the analytical processes and the conclusions obtained, with at most 8 pages:

• Heading 1: Arial, Size 12 pt, in bold

• Heading 2 (if needed): Arial, Size 11 pt, in bold and italic

• Text: Arial, Size 10 pt, line space of 1.5 points.

• Margins: The default ones in word (Top, Bottom, Left and Right as 1”).

All the figures and tables should be included in the Annexes (at the end of the document) and referenced in the body text, and are not included on those 8 pages mentioned previously. The reports that do not follow the specified conditions will suffer penalizations on the grade. The file naming format should be ”202122 Cluster GroupXX Report.pdf”, where ”GroupXX” should be your group number.

Group Project

BookMe

DATA SCIENCE AND MACHINE LEARNING 2022

March 1, 2022

**1 Business problem**

BookMe is a well-established international company, working on the hospitality sector. It provides accommodations to tourists and travelers, whether for leisure or business purposes.

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.

The profit growth perspectives for the next three years are fickle, so the marketing team needs a new efficiency program to improve their activities and boost their campaigns, in order to invert the situation. Our main goal as the analytical team in charge of this project, was to create different segmentations for the BookMe clients and services and propose targeted marketing campaigns.

**2 Data**

In order to suggest a new marketing plan, we have explored the available users’ dataset[[1]](#footnote-1), provided by the BookMe company and through the usage of several analytical techniques, we have applied clustering methods to find patterns in client’s behavior.

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

Our first step was to import the available data[[2]](#footnote-2) and remove the duplicate entries[[3]](#footnote-3). Our final dataset was composed by a total of 15,586 rows/observations and 21 columns/attributes[[4]](#footnote-4).

**3 Exploration**

We have started the analysis with a statistical exploration over the available data[[5]](#footnote-5). Based on that, we were able to get some insights over the BookMe customers and identify incoherencies and issues in our data.

Regarding the satisfaction level segmentation, we were able to conclude the following. 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.

Regarding the customer segmentation, we were able to conclude as following. Around 80% of the customers are registered in the platform for about 1 year, at least, and more than 50% of them are in a ‘nochurn’ situation, which gives us 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 48% of preference. When combined with the “TravelType”, we found a pattern in which the ‘single’ and ‘suite’ room types were mostly booked by travelers in business and the ‘double’ ones were the preference for leisure travels[[6]](#footnote-6).

Using some visualization techniques, we discovered that the most recent clients, meaning the ones registered in the BookMe platform for less than 1 year, were the ones with more ‘churn’ than ‘nochurn’ situations[[7]](#footnote-7). This was, obviously, worth more exploration, because it could be important for the marketing team to create new incentives for newcomers, in order to invert this situation.

In terms of issues, we have identified 195 missing values in the ‘Year\_Birth’ variable, 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, this survey is only provided to registered users, after each stay. For that reason, we do not believe that these values should be acceptable. Also, the maximum rate value for “Wifi” was appointed as ‘6’. This was clearly an incoherence, because, according with the business rules, the possible rate scale is between ‘0’ and ‘5’. The last incoherence we have identified was related with the variable “Longevity”, which had three possible values, instead of two: ‘yes, ‘no’ and ‘y’. Clearly, ‘yes’ and ‘y’ correspond to the same reference.

4 Preprocess

Our next step was to preprocess our data, which means find a solution for the issues and incoherencies, identified previously.

In terms of shape and density, 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[[8]](#footnote-8) values, related with the variables “Amenities”, “Staff”, “PriceQuality”, “CheckOut” and “Cleanliness”, as per the kurtosis[[9]](#footnote-9) calculations, we were not able to identify the presence of any potential outliers. Even though, we still have explored this possibility, using some visual techniques[[10]](#footnote-10). 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 our data. As per the rule of thumb, no more than 3% of the outliers should be removed, so we have kept all the registers, since we had no way of differentiating them.

Regarding missing values identified previously in the ‘Year\_Birth’ variable, we have also replaced its incoherencies with missing values, ending up with a total of 1,246 entries to fill in[[11]](#footnote-11). Even though this variable was not highly correlated with any other, we still have tried to use the KNNImputer technique, to fill in the missing values. After comparing the descriptive statistics of the variable ‘Year\_Birth’ distribution, before and after the imputation, we concluded that the result was good, because the main values have pretty much remained equal, and the standard deviation decreased a little[[12]](#footnote-12).

Regarding misclassifications identified in the variable ‘Wifi, we have replaced the 36 entries with a rate of ‘6.0’, by the maximum possible value of ‘5.0’[[13]](#footnote-13).

Afterwards, we have created new variables, based on the available attributes. We have noticed that the variable “Name” could be converted into a new attribute “Gender”[[14]](#footnote-14), based on the available prefixes: ‘Mr’ and ‘Ms’. We have also converted the ‘Year\_Birth’ column into ‘Age’[[15]](#footnote-15). Then, we have created a new variable called ‘satisfaction\_avg’, to assess the average satisfaction level of each customer with the different services, provided by the BookMe company.

Finally, based on a heatmap[[16]](#footnote-16) figure, we were able to identify which features were more correlated with one another. The result was as following: “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). Basically, the most correlated features correspond to BookMe services, so we had the idea to combine the 14 different types into groups and calculate the corresponding average satisfaction, per client. With this in mind, we have created four new attributes: ‘accomodation\_avg’[[17]](#footnote-17), which combines the variables ‘Comfort’, ‘Amenities’, ‘RoomSpace’, ‘Wifi’ and ‘Cleanliness’; ‘reception\_avg’[[18]](#footnote-18), which combines the variables ‘ReceptionSchedule’, ‘Staff’, ‘Checkin’, ‘OnlineBooking’ and ‘CheckOut’; ‘catering\_avg’[[19]](#footnote-19), which combines the variables ‘FoodDrink’ and ‘BarService’ and ‘key\_factors\_avg’[[20]](#footnote-20), which combines the variables ‘Location’ and ‘PriceQuality’.

Next, we have performed some binning over the ‘RewardPoints’ variable and converted it into a new one, called ‘RewardPoints\_bins’[[21]](#footnote-21). This way, the original values were combined into 3 different groups – ‘low’, ‘medium’ and ‘high’ – all with the same equal-width bins.

Our last transformation was performed over the variable ‘Longevity’, in order to reclassify the incorrect values identified previously[[22]](#footnote-22).

Concerning data reduction, due to the multicollinearity problem, we dropped some highly correlated variables, like the ‘Name’, which was replaced by ‘Gender’ and the ‘Year\_Birth’, converted into ‘Age’.

To ensure that our data had only numerical values, we have created dummy variables and finally we were able to scale the entire dataset, in order to apply the required algorithms for the clustering exercise[[23]](#footnote-23).

**X Annexes**

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 3 - 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 - Skweness values

Uma imagem com texto

Descrição gerada automaticamente

Figure 7 - Kurtosis values

Uma imagem com texto, recibo

Descrição gerada automaticamente

Figure 8 – Descriptive statistics for categorical variables

Graphical user interface, application

Description automatically generated

Figure 9 – Heatmap

A screenshot of a computer

Description automatically generated with medium confidence

Figure 10 - Countplot of "Longevity" and "Churn"

Chart, bar chart

Description automatically generated

Figure 11 - Countplot of "RoomType" and "TypeTravel"

Chart, bar chart

Description automatically generated

Figure 12 - Boxplots for 'Amenities' and 'Staff'

Chart, box and whisker chart

Description automatically generated

Figure 13 - Boxplots for 'PriceQuality' and 'CheckOut'

Chart, box and whisker chart

Description automatically generated

Figure 14 - Boxplot for 'Cleanliness'

Chart

Description automatically generated

Figure 15 - Missing values distribution (before)

A picture containing text

Description automatically generated

We have replaced the entries with ‘Year\_Birth’ higher than ‘2004’ by missing values, in addition to the original ones.

Figure 16 - Missing values distribution (after)

Text

Description automatically generated with medium confidence

Figure 17 - 'Year\_Birth' descriptive statistics before imputation

Text

Description automatically generated

Figure 18 - 'Year\_Birth' descriptive statistics after imputation

Text

Description automatically generated

Figure 19 - 'Wifi' values before correction

Table

Description automatically generated

Figure 20 - 'Wifi' values after correction

Table

Description automatically generated

Figure 21 - Distribution of new variable 'Gender'

Text

Description automatically generated with medium confidence

Figure 22 - Distribution of new variable 'Age'

Text, letter

Description automatically generated

Figure 23 - Distribution of new variable 'satisfaction\_avg'

Text, letter

Description automatically generated

Figure 24 - Distribution of new variable 'accomodation\_avg'

Text, letter

Description automatically generated

Figure 25 - Distribution of new variable 'reception\_avg'

Text, letter

Description automatically generated

Figure 26 - Distribution of new variable 'catering\_avg'

Text

Description automatically generated

Figure 27 - Distribution of new variable 'key\_factors\_avg'

Text, letter

Description automatically generated

Figure 28 - Distribution of new variable 'RewardPoints\_bins'

Text

Description automatically generated with medium confidence

Figure 29 - 'Longevity' values before transformation

Text

Description automatically generated

Figure 30 - 'Longevity' values after transformation



Figure 31 - First 5 entries of final dataset

Table

Description automatically generated

1. Please, refer to Figure 1 - Summary of dataset’s Attributes and Description. [↑](#footnote-ref-1)
2. Please, refer to Figure 2 - First three entries of BookMe dataset. [↑](#footnote-ref-2)
3. Please, refer to Figure 3 - Duplicated entries removed. [↑](#footnote-ref-3)
4. Please, refer to Figure 4 - Information on BookMe dataset. [↑](#footnote-ref-4)
5. Please, refer to Figure 5 - Descriptive statistics for numerical variables and Figure 8 – Descriptive statistics for categorical variables. [↑](#footnote-ref-5)
6. Please, refer to Figure 11 - Countplot of "RoomType" and "TypeTravel". [↑](#footnote-ref-6)
7. Please, refer to Figure 10 - Countplot of "Longevity" and "Churn". [↑](#footnote-ref-7)
8. Please, refer to Figure 6 - Skweness values. [↑](#footnote-ref-8)
9. Please, refer to Figure 7 - Kurtosis values. [↑](#footnote-ref-9)
10. Please, refer to Figure 12 - Boxplots for 'Amenities' and 'Staff', Figure 13 - Boxplots for 'PriceQuality' and 'CheckOut' and Figure 14 - Boxplot for 'Cleanliness'. [↑](#footnote-ref-10)
11. Please, refer to Figure 15 - Missing values distribution (before) and Figure 16 - Missing values distribution (after). [↑](#footnote-ref-11)
12. Please, refer to Figure 17 - 'Year\_Birth' descriptive statistics before imputation and Figure 18 - 'Year\_Birth' descriptive statistics after imputation. [↑](#footnote-ref-12)
13. Please, refer to Figure 19 - 'Wifi' values before correction and Figure 20 - 'Wifi' values after correction. [↑](#footnote-ref-13)
14. Please, refer to Figure 21 - Distribution of new variable 'Gender'. [↑](#footnote-ref-14)
15. Please, refer to Figure 22 - Distribution of new variable 'Age'. [↑](#footnote-ref-15)
16. Please, refer to Figure 9 – Heatmap. [↑](#footnote-ref-16)
17. Please, refer to Figure 24 - Distribution of new variable 'accomodation\_avg'. [↑](#footnote-ref-17)
18. Please, refer to Figure 25 - Distribution of new variable 'reception\_avg'. [↑](#footnote-ref-18)
19. Please, refer to Figure 26 - Distribution of new variable 'catering\_avg'. [↑](#footnote-ref-19)
20. Please, refer to Figure 27 - Distribution of new variable 'key\_factors\_avg'. [↑](#footnote-ref-20)
21. Please, refer to Figure 28 - Distribution of new variable 'RewardPoints\_bins'. [↑](#footnote-ref-21)
22. Please, refer to Figure 29 - 'Longevity' values before transformation and Figure 30 - 'Longevity' values after transformation. [↑](#footnote-ref-22)
23. Please, refer to Figure 31 - First 5 entries of final dataset. [↑](#footnote-ref-23)