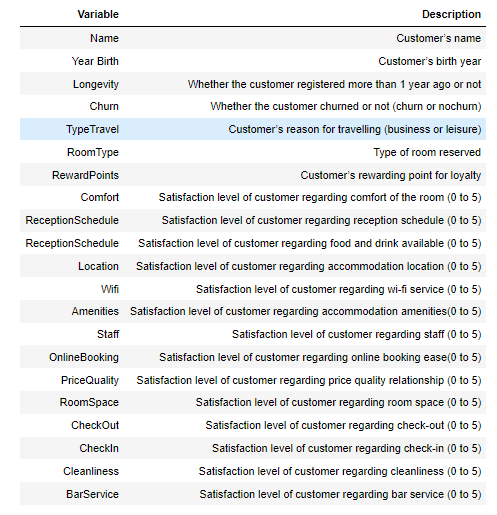
**Business Description**

Descrever o tipo de dados o que estes são, por exemplo:

“Estes dados são de clientes de um hotel que avaliaram o serviço”

1. **Data Set**

The first step is analyze the data ser that we have and be aware of all of features and kind of variables. So, the data set has 15 589 rows and 21 columns. Below we can see the basic description of meaning of each variable.

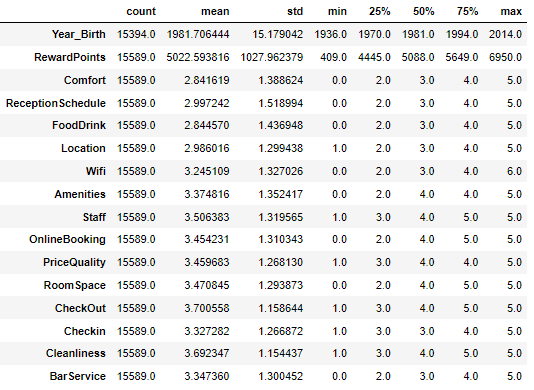


1. **DataFrame Information and description/Undestanding**

Describing the type of variables of this data frame, we have five o*bjects* (Churn, Name, Longevity, TypeTravel and RoomType), one *float* (Year\_Birth ) and fifteen *intengers* (RewardPoints, Comfort, ReceptionSchedule, FoodDrink, Location, Wifi, Amenities, Staff, OnlineBooking, PriceQuality, RoomSpace, CheckOut, CheckIn, Cleanliness and BarService).

The objects are the categorical variables the float and intengers are numerical variables.

**About the numerical variables we present below the descriptive statistics.**



About these statistics we are able to see the following:

* Almost all variables left between 0 and 5
* Location, Staff, PriceQuality, CheckOut, CheckIn, Cleanliness doesn’t have any satisfaction level of 0
* The **age of the customers** in in this data set go **from 8 years old** to **86 years old** (if we consider the year of 2022)
* The **average age** of customers in this data frame is **41 years old** (if we consider the year of 2022)
* There are **missing values** on **Year\_birth variable**
* In average the **CheckOut** gets the **higher** satisfaction level between the customers
* In average the **Comfort** gets the **lower** satisfaction level between the customers
* The WiFi variable **has a maximum of 6**
* The variables has different scales

**About the categorical variables, we can see below what is the expected values in each of these variables:**

* Churn
  + “Churn”
  + “Nochurn”
* Name
  + Name of customers
* Longevity
  + “Yes”
  + “No”
  + “Y”
* TypeTravel
  + Business
  + Leisure
* RoomType
  + Single
  + Double
  + Suite

**Skewness**

Skewness is a measure of how balanced the distribution is. It measures the symmetry of the distribution. A balanced distribution, such as the normal distribution, has a skewness value of zero.



After analysing the skewness of all numerical variables, we can saw that:

* **Positive Skewness/Left Tail**

The variables have positive skewness that means that have a left tail distribution (mode<median<mean) and we have no cases in this report.

* **Zero Skewness/Symetric Distribution**

The variables have no skewness or almost no skewness (between -0.5 e 0.5) that means that have symmetric distributions: **“Year\_Birth”,** **“Location”, “RewardPoints”, “Comfort”, “ReceptionSchedule”, “FoodDrink”, “Wifi”, “Checkin” and “BarService”.**

* **Negative Skewness/Right Tail**

The variables have negative skewness, only moderated skewness (between -0.5 and -1) that means that have a right tail distribution (mode>median>mean) and we have some cases in this report: **“Amenities”, “Staff”, “OnlineBooking”, “PriceQuality”, “RoomSpace”, “CheckOut”, “Cleanliness”.**

In terms of business understanding this tell us that the most of numeric variables, that are evaluations from the clients, are **negative skewness which means that the most of evaluations are better that the half of available scale**. The “location” is the only one with which have symmetric distribution, so has **more evaluations closer to the half of scale**.

**Adicionar no final os plots todos**

**KURTOSIS????**

**Correlation**

Fazer depois do congruence check e criação de novas variáveis

1. **Data Preparation**

With all data understanding we become aware of, we need to do prepare all data to start modeling. In this data preparation we did the following procedures:

* 1. Fulfil the missing values
  2. Outlier observation and removing
  3. Straight lines removing
  4. Misclassifications and Reclassifications
  5. News features
  6. Dummy and Binary variables
  7. Data normalization
  8. Correlation Check

3.1 Fulfil the missing values

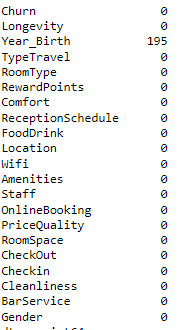
Most machine learning algorithms need complete datasets without missing values, however as we can see in descriptive statistics above, the “year\_Birth” variable has some missing values. We’ve checked it with “isna()” command and we doublecheck what we have saw before, “Year\_Birth” have 195 missing values.

There are several ways to fill these values but the KNNimputer algorithm seems the best option. The KNNimputer is used to predict missing values based on data points that we already have in dataset. The value “k” refers to the number of neighbours needed to forecast unseen data points. Those neighbours will be at certain distance from missing points and the computation of that is very important to of this algorithm.

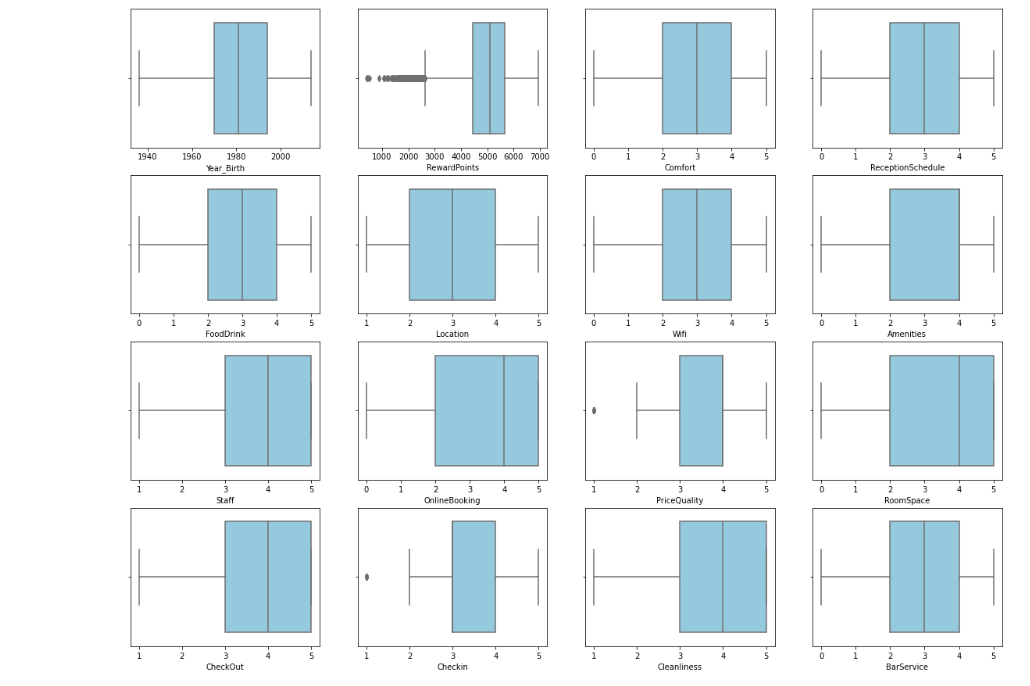
“k” choose – WHY 2? – Quanto maior menor o erro

Although there is no theoretical approach to choose the value of “k”, it is common practice to evaluate the performance of the model based on the errors of some scoring test. As we increase the value of “k”, it is possible to evaluate the difference between the scoring test and choose the value of “k” associated with the lowest value of the error. Usually, as the number of neighbours increases, the error tends to stabilize, meaning the model reached convergence.

EXPLICAR COMO USAMOS E QUAL FOI O K



3.2 Outlier observation and removing



We use the boxplot to check the ouliers and 3 possible options came up: “RewardPoints”, “PriceQuality” and “CheckIn”.

In terms of “RewardPoints” variable we calculate the interquartile range (IQR) which is calculated subtracting the first quartile to the third quartile. The output of this calculation is the lower and the higher bound of IQR. The lower bound is calculated by subtracting to the first quartile to 1.5 times the IQR values. The higher bound is calculated by adding to the third quartile to 1.5 times the IQR values.

With the bounds calculated and as we can see in “RewardPoints” box plot the outliers and below the lower limit of box plot, so, to remove the outliers we decide to maintain the values that are above the lower IQR bound. 193 outliers were removed.

After some research and analysis on the subject, we decided not to remove outliers other two ratings variables that presented outliers in the boxplot visualization. Those are 'PriceQuality' and 'Checkin'. The reason we did not remove them was because they were too many entries that would have to be removed, and if we did, we would effectively be shortening the rating scale for the rating data.

3.3 Straight lines removing

Instead, we will look for straight lining on the ratings entries, meaning people who answered all the answers with the same variable, which can mean they were in a rush and decided to fill the survey as fast as they could.

There were only to entries with exactly the same answers, then we remove it.

3.4 Misclassifications and Reclassifications

Some categorical variables have a misclassification that need to be fixed in order to user it without any misunderstanding.

The “Longevity” has three different values, “No”, Yes” and “Y”, we decide do transform the “y” in “yes” answers and keep only the “yes” and “no” answers that makes all sense.

In terms of numerical variables, we find a misclassification as well. The “Wifi” variable has a maximum value of 6, once the upper bound of scale is 5 this is a misclassification. We transform that “6” value in the maximum value of scale, which is “5”.

3.5 News features

In order to facilitate the data understanding we create two more features according the variables that we already had. One of them is the gender, this variable is included in the first three characters of “name” variable, so the decide to create a new binomial variable based on this first characters. The male ones, which has “Mr.” in first characters, we attribute the value of “1” in “gender” variable, for the other ones (only “Ms.”) we attribute the value of “0”. The “gender” binomial variable is created. After create this new variable we decide to drop the “name” variable.

Other variable that we create to better understanding of data is the “age” variable. This variable was based on “year\_birth”, we calculated the age of each customer with the year of analysis on this data frame (2022) and keep it in a new numerical variable. After create this new variable we decide to drop the “year\_birth” variable.

3.6 Dummy and Binary variables

After dropping some categorical variables once we extract the information that we need from them, we still having some categorical variables that we need to transform in numerical variable. To be useful to use in Kmeans algorithm, we cannot use categorial variables but we need to transform it in numerical and binary variables.

Churn

The “Churn” we transform in “1” every answer that represent a “churn” customer and “0” every customer that represent a “nochurn” customer.

Longevity

After reclassifying this variable and keep onlye the “yes” or “no” answers we transform this variable where the “yes” correspond to “1” value and the “no” correspond to “0” value.

TypeTravel

Here we have two answers as well, the “business” and the “leisure” which we transform in “1” and “0” correspondingly.

RoomType

In this variable the case is little different, we have three different answers, “single”, “double” and “suite”. In order to get numerical variables from this one we transform in two variables: “RoomType\_single” and “RoomType\_suite”.

The “RoomType\_single” means that is a single room when the value is “1” and a double room when the value is “0”. The “RoomType\_suite” means that is a suite when the value is “1” otherwise is not a suite room.

3.7 Data normalization