1. Data understanding

First, was done some statistics on the variables to better understand their behaviour, in particular, of the continuous variables. Looking to the visualizations produced and summary statistics was possible to take the following conclusions:

Age – Given the observed histogram and that mean ≈ median we can conclude that this variable presents an approximation to a normal distribution. Consequently, the dataset presents a bigger concentration of individuals with ages values around the 46/47 years old, with smaller concentrations of individuals with smaller and bigger ages. It's also possible to observe ages smaller than 0 and bigger than 100, that might be errors in the imputation.

AverageLeadTime – It's possible to observe that most of the individuals do the bookings near the date of accommodation, however it's possible to observe individuals that do bookings almost 2 years before the staying. At least 25% of the population does the booking in the same day of the accommodation. It's present at least one negative value in this variable, this might be errors in the imputation.

BookingsCanceled - Most of the population has done no cancelations at all (at least 75%). There's no individual that has cancelled the booking more than 15 times.

BookingsNoShowed - Most of the population has done no "no show" at all (at least 75%). There's no individual that has "no showed" the booking more than 3 times.

BookingsCheckedIn - At least 25% of the clients haven't done at least 1 check in, that might be an error in the data or people that were replaced due to overbooking (if other variables have values). Most of the clients have a low number of checked ins. The maximum number of checked ins of individuals that we have is 75.

DaysSinceCreation - All the clients were created for at least 30 days, having clients resisted for almost 4 years. Almost 50% of the population is with us for less than 500 days, having a bit more than 50% with us from 500-1385 days.

LodgingRevenue - At least 25% of the population gives no lodging revenue at all, however this could be explained by the fact that at least 25% of the clients haven't done at least 1 check in. Only 25% of the clients give the bigger revenues of 393.3-21781 €, being most of the population concentrated in smaller revenues.

OtherRevenue - At least 75% of the clients don't give in other revenues values higher than 84. However, there are clients that give in other revenues 8859.25€. Again, it's seen that at least 25% of the clients' don´t give any other revenue at all. This might be explained by the 25% of the clients with no check in.

PersonNights - (Again it's seen that at least 25% of the clients with 0 PersonNights that might be explained by the 25% of the clients with no check in). At least 75% of the population don't have a bigger value than 6, however there are clients with 116 PersonNights.

RoomNights - (Again it's seen that at least 25% of the clients with 0 PersonNights that might be explained by the 25% of the clients with no check in). At least 75% of the population don't have a bigger value than 3 however there are clients with 185 PersonNights.

Special Requests - We can see that at least 25% of the clients' requisite a King-Sized bed, being this the most demanded preference. Besides that, seems that only 'SRTwinBed', 'SRQuietRoom', 'SRCrib' and 'SRHighFloor' have a significant number of requests.

2. Coherence / Discrepancy Verification (Data cleaning)¶

(Hash Part)

A unique person is given by the combination of his name and document ID (**NameHash** and **DocIDHash**). Said that, doing this validation, were found a total of 5004 records that had this combination more that one time, ie, the same person had more than one record. As a solution, these records were grouped by **NameHash** and **DocIDHash:**

- summing the **revenues**, the **bookings**, **PersonNights** and **Room Nights**;
- keeping the maximum number of **DaysSinceCreation**;
- doing the average of **AverageLeadTime**
- doing the mode of the remaining variables

At the end the 5004 records were transformed into 2175 unique persons' records. **NameHash** and **DocIDHash** columns were dropped since they don´t bring any value to our analysis.

3. Feature engineering

3.1 Create new calculated features

In order to improve the performance of the machine learning algorithm applied and to improve the insights obtained in the clustering, were designed the next new features:

'Age_bin' – However the age variable looks normal distributed, by binning this variable a set of patterns are found in this continuous variable which are easy to analyse and interpret.

- '**DistributionChannel**' and **'Age'** dropped and transformed to **'DistributionChannel'** and **'Age_bin'** as dummy variables to further consider them in our analysis.

'AvgRevenueYear' – average of the total spending by the customer per year – allows to compare among clients the revenue generated by them considering the time that they are our customer.

- **LodgingRevenue** and **OtherRevenue** dropped, for this analysis is enough to look at the total revenue (summing those variables) and obtain **'AvgRevenueYear'.**

'Personnightsperbook' – Average of personnights that the client does by booking – allows to compare the variable PersonNights among the clients.

- **PersonNights** dropped**.**

'Revenueperroomnight' – Average Revenue (total spending by the customer) generated by RoomNight – allows to understand how valuable is an extra RoomNight for each customer.

- **RoomNights** dropped.

'BookingsNoShowedprt' / **'BookingsCheckedInprt' / 'BookingsCanceledprt'** – Percentage of bookings the customer made but subsequently made a "no-show"/ Check In / Cancellation – allows to compare those variables among customers.

- **'BookingsNoShowed'**, **'BookingsCheckedIn'** and **'BookingsCanceledprt'** dropped.

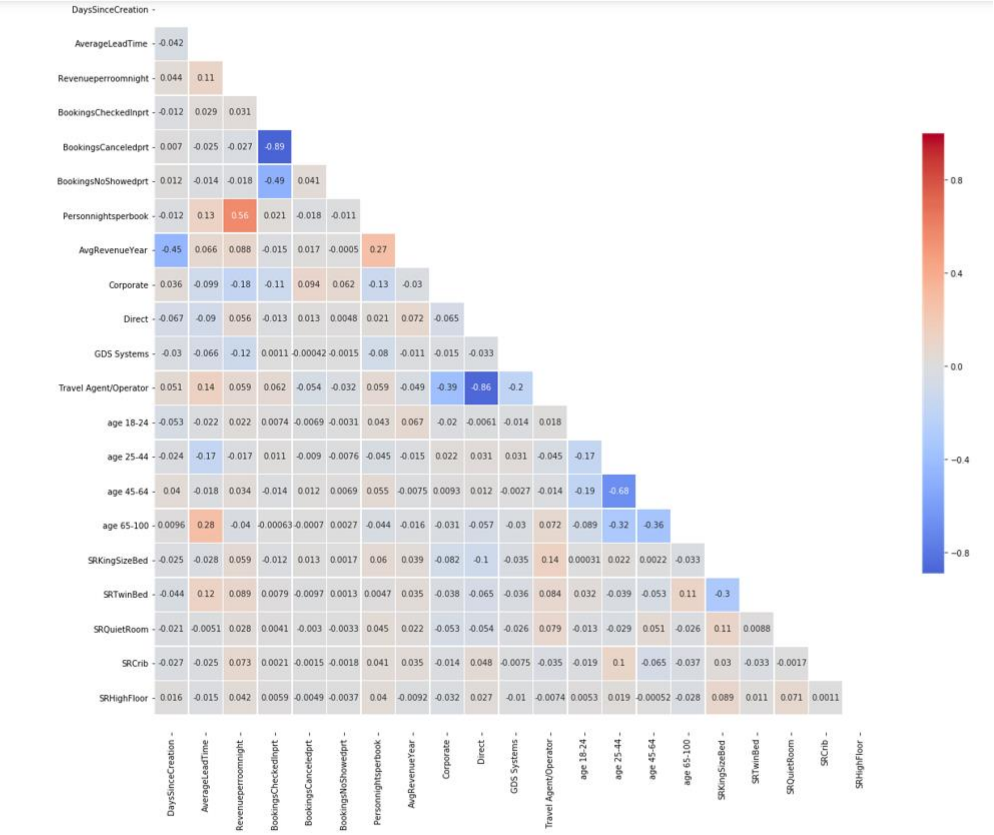
3.2 Import external data source to complement the analysis

Was added to the data frame the **region** and **sub-region** of the client, provided by an external source, to further produce some visualizations and understand the impact of those variables.

3.3 Input space reduction

First, concerning the special requisites, we only wanted to consider the ones that were selected by a significant part of the population. This way, we just kept the top 5 preferences: 'SRKingSizeBed', 'SRTwinBed','SRQuietRoom','SRCrib' and 'SRHighFloor.

Next, the correlation were plotted in order to avoid redundancy in the dataset:



BookingsCheckedInprt ~ BookingsNoShowedprt (-0.49)

BookingsNoShowedprt has only 31 values different than zero, this variable doesn't bring the needed value for our analysis due to its low representation. For those reasons this variable was dropped.

BookingsCheckedInprt ~ BookingsCanceledprt (-0.89)

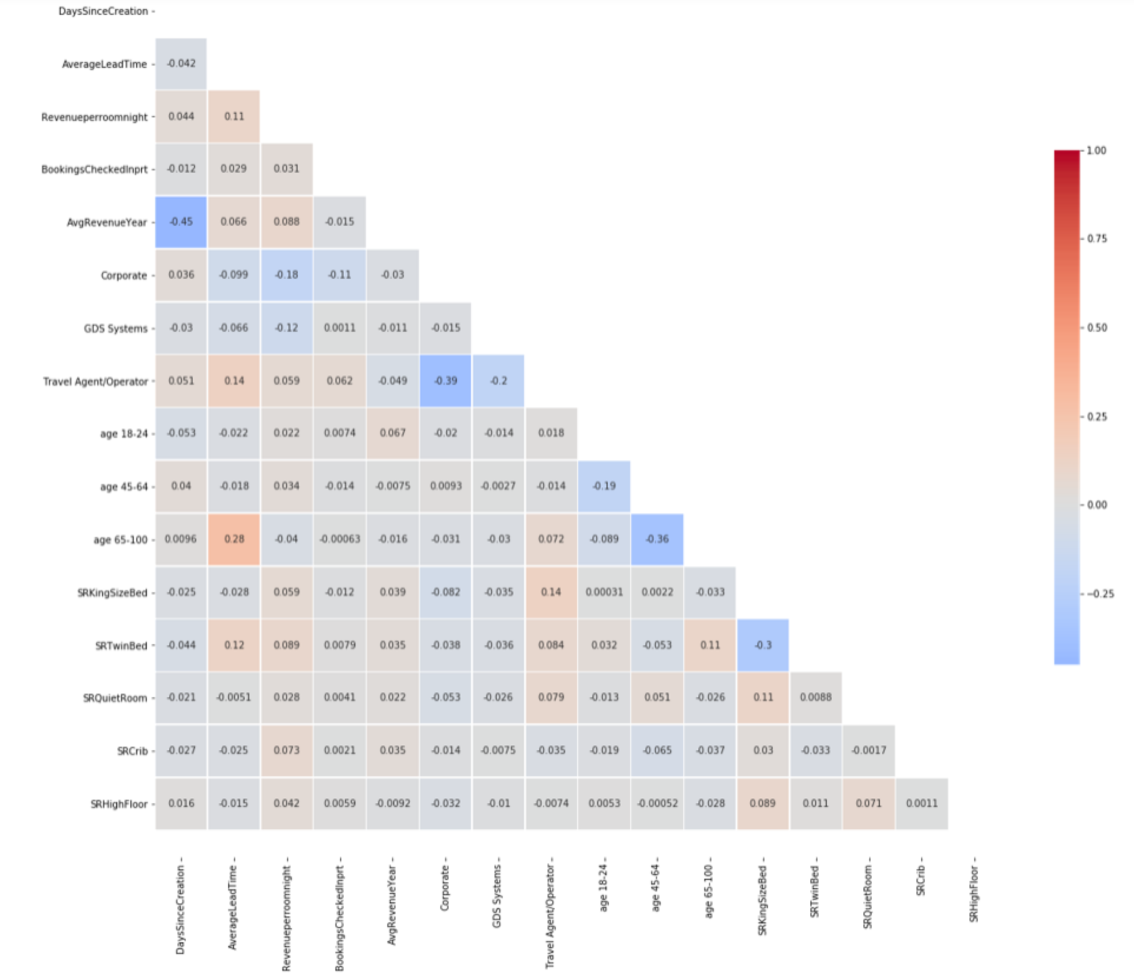
This high correlation was explained by the low representation of BookingsNoShowedprt. Excepting 31 cases, BookingsCheckedInprt = 100 – BookingsCanceledprt therefore, BookingsCheckedInprt was kept.

PersonNightsperbook ~ Revenueperroomnight (0.56)

These two variables were related and, given their correlation, not relevant with each other. For the analysis was considered that **Revenueperroomnight** had the biggest value, therefore **PersonNightsperbook** was dropped.

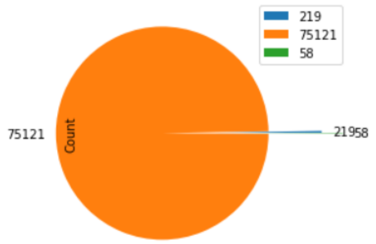
Travel Agent/Operator ~ Direct (-0.86) and age 45-64 ~ age 25-44 (-0.68)

Since they are dummy variables concerning the **DistributionChannel (Age_bin)**, instead of having n dummy variables for the n categories, we kept n-1 dummies by dropping the **Direct** (**age 25-44**) column that can be decoded by the others.



5. Data visualization

5.1 Combine features on different plots to find more insights about the business

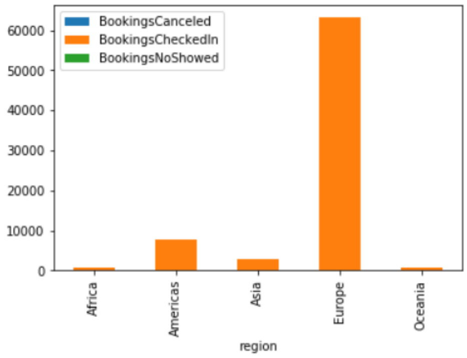


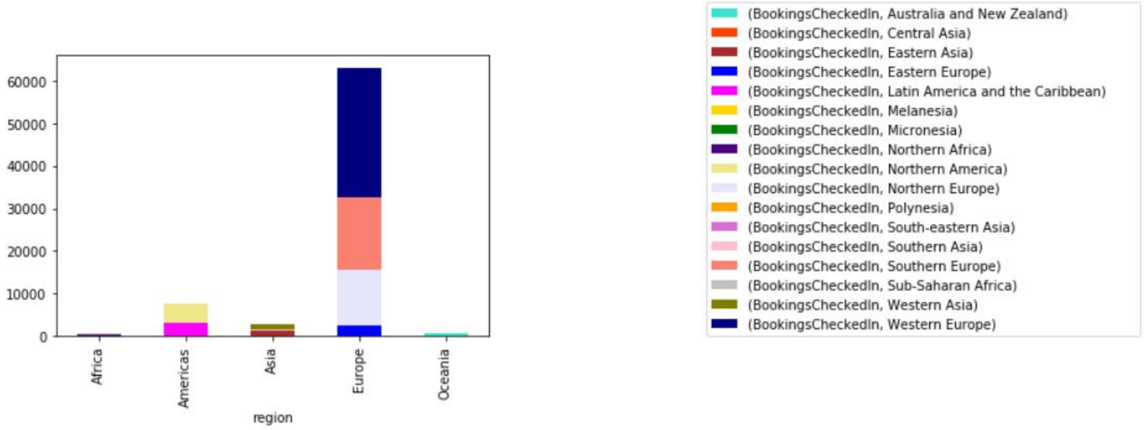
After the data coherence / discrepancy verification we have, in our database, a total of 75121 bookings checked in, 219 bookings cancelled and 58 “no showed”.

Most of our checked in clients are represented by Europe, followed by America and Asia.

Considering the sub-regions, we can observe that:

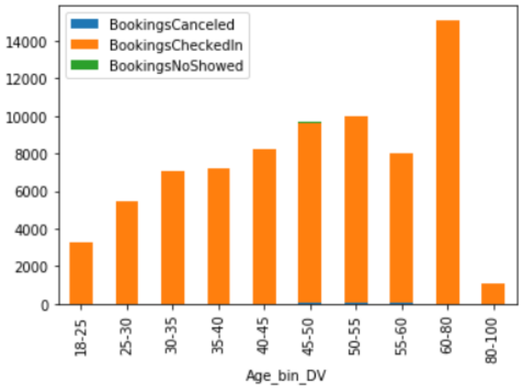
- Our Europeans clients are mostly from the Wester Europe, followed by Southern and Northern Europe.
- Our Asian clients are mostly from Eastern and Western Asia
- Our American clients are almost even distributed by Latin America and Caribbean and Northern America.

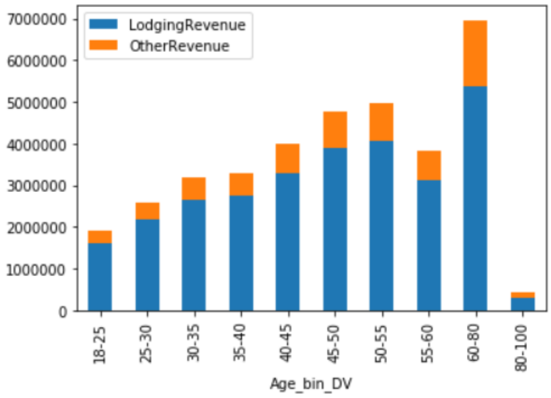




Looking for the ages, most of the checked in clients have ages of 60-80. Client with ages of 18-25 have the lower representation in check ins.

By consequence, the same distribution of the population by ages is observed when comparing the age bins with revenue. The Lodging Revenue has the bigger representation on total revenue by age bin too.





5.2 Conclude about the best variables to do the clustering

Thus, after processing 28 variables have been considered for further analysis. Those variables are DaysSinceCreation, AverageLeadTime, RevenuePerRoomNight, BookingCheckedIn, AvgRevenueYear, Corporate, GDP Systems, Travel/Agent Operator, Age bins and 5 Special requests.

The 5 Special requests chosen account to more than 98% of records so the rest of the SR have been disregarded for their irrelevance.

The nationalities won't be considered since that Europe has much bigger presence than other regions.

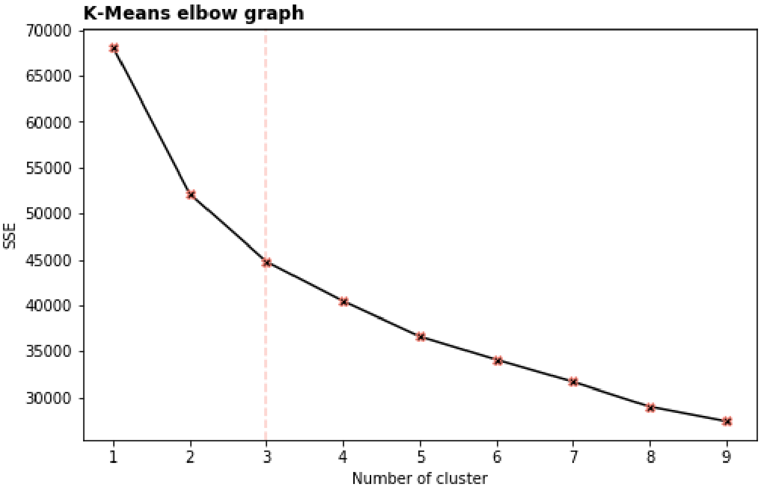
The age bins will be considered; however, the 80-100 bin was added to the 60-80 bin due to its low number of records.

Principle Component Analysis

In order to achieve the best customer clustering, we use PCA, a dimensionality reduction technique Before implementing PCA we normalized the Data with StandardScalar module. StandardScalar scales to the unit variance after subtracting the mean of each feature, enabling us to apply PCA(https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html).

As an output of PCA we receive 15 clusters that cumulatively explain 98.18 % of variance. INSERT_PCA_GRAPH (I think).

Subsequently, based on the newly acquired 15 principle components we use the elbow method(https://www.scikit-



yb.org/en/latest/api/cluster/elbow.html)to define the optimal number of K clusters to be used in the final clustering algorithm. As we can observe from the Sum of Squared Errors graph, 3 is suggested as the cut off point for the number of clusters.