BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ALAYTICS

Hotel Customer Segmentation

Euclidean Consultancy Group

Ernesto Madrid, number: m20190559

Filipa Cerqueira, number: r2016677

Laura Castro, number:m20190269

Norayr Meliksetyan, number: m20190687

March 2020

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION

Due to the evolution of digital transformation of business models and the available data, during the last years, companies prefer to support their decisions on data-driven strategies. In this report, a Hotel Customers Segmentation Business Case is developed to precisely close the gap between customers data and the realignment of the company strategies.

The Business Case is addressed following the CRISP-DM methodology to organize (Ncr & Clinton, 2000) and develop the life cycle of this project, in which Principal Components Analysis (PCA) and K-means clustering were performed on the Modelling phase in order to achieve the customers segmentation.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

The Hotel chain C has provided to this consultancy group a dataset with 111.733 customers bookings records aiming to deliver to the marketing manager a robust customer segmentation based on the available features of the mentioned dataset. During previous years, the customer segmentation used by this Hotel chain was based only in sales origin, but not in demographics or behavioral characteristic.

2.2. OBJECTIVES

The actual Business Case aims to change the customer segmentation of the Hotel chain C, provide new strategy paths to for the marketing department, and with those strategies, increase the revenues by taking advantage of the customers preferences.

The main advantage of applying Hotel Customer Segmentation is to know the specificity of each customer group. The more specific the market segmentation, the easier it is to allocate investments and strategies to find the correct customers to generate higher revenues (Waida, n.d.). Nevertheless, most of the cases customers hide motivations of purchase behavior and satisfaction, therefore exploring available data and performing Customer Segmentation Modelling is so valuable (Rondan-Cataluña & Rosa-Diaz, 2014).

2.3. SUCCESS CRITERIA

In order to develop success criteria of any project it is crucial to have a comprehensive understanding of the business processes, financial statements and strategic goals. However, such that there is relatively limited amount of information about the company in question, we defined the success criteria as such:

- Develop a well-functioning model for current and future customer segmentation needs of the company
- Provide a new marketing plan overview based on the customer segmentations

Based on these points, the

3. PREDICTIVE ANALYTICS PROCESS

3.1. DATA UNDERSTANDING

Initially, a summary statistic was executed on the variables to better understand their behaviour, in particular, the continuous variables. Looking at the visualizations, and the summary statistic, it was possible to make the following descriptive analysis over each provided feature:

- **Age**: Given the observed histogram, it is possible to say that the mean is almost equal to the median, presenting an approximation to a normal distribution. Consequently, the dataset presents a higher density of individuals with ages around the 46 years old, with smaller density 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 calculation or imputation.
- **AverageLeadTime**: It is possible to observe that most of the individuals do the bookings near the date of accommodation, however a few individuals that do bookings almost 2 years before the staying. At least 25% of the population do the booking in the same day of the arrival. It is 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 is 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 is no individual that has "no showed" the booking more than 3 times.
- **BookingsCheckedIn**: At least 25% of the clients have not 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 is 75.
- **DaysSinceCreation**: All the customers were created for at least 30 days, having customers registered for almost 4 years. Almost 50% of the population is with us for less than 500 days, having a bit more than 50% with the Hotel company from 500 to 1.385 days.
- **LodgingRevenue**: At least 25% of the population give no lodging revenue at all, however this could be explained by the fact that at least 25% of the customers have no check in. By other hand, only 25% of are between 393,30 and 21.781,00 €, being most of the population concentrated in smaller revenues.
- **OtherRevenue**: At least 75% of the customers do not provide more than 84,00 € on this kind of revenue. However, a few customers give 8.800,00€. As in *LodgingRevenue*, it is seen that at least 25% of the customers do not contribute at all, this might be explained because 25% has no check in.
- **PersonNights**: At least 75% of the population do not have a value bigger than 6, however there are customers with 116 *PersonNights*. The rest 25% do not has a check in value.
- **RoomNights**: At least 75% of the population do not have a value bigger than 3 however there are customers with 185 *PersonNights*.
- **Special Requests**: At least 25% of the customers requested a King-Sized bed, being this the most demanded preference, followed by '*SRTwinBed*', '*SRQuietRoom*', '*SRCrib*' and '*SRHighFloor*'.

3.2. DATA PREPARATION

3.2.1. COHERENCE / DISCREPANCY VERIFICATION (DATA CLEANING)

In this section is verified that all the values for the variables were in fact possible considering what each variable is measuring. Using the variables summary statistics, a general validation was done, conforming some variables did not have discrepancies. The ones with incorrect values were examined one by one:

- **Nationality**: A dataset containing all nationalities in the world was compared to the nationalities of the clients and is was confirmed that all customers had, indeed, valid nationalities.
- **Age**: All customers with an age below 18 or above 100 were excluded. The values for age restriction were chosen considering that people younger than 18 cannot book a hotel room, as it is the minimum legal age. The maximum age was set at 100, considering the human life expectancy.
- **AverageLeadTime**: Negative values were excluded.
- **MarketSegment**: All values were confirmed.
- **NameHash and DocIDHash**: A unique person is given by the combination of his name and document ID (*NameHash* and *DocIDHash*), but 5.004 records had this combination more than one time, this is, 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 *RoomNights*
  - keeping the maximum values on *DaysSinceCreation*
  - calculating the average of *AverageLeadTime*
  - calculating the mode of the rest of the variables

At the end the 5.004 records were transformed into 2.175 unique persons' records. *NameHash* and *DocIDHash* columns were dropped since they do not bring any value to our analysis.

3.2.2. FEATURE ENGINEERING

3.2.2.1. Creating new calculated features

In order to improve the model performance and combine variables to get more insights from the models results, the following features were designed:

- **Age_bin**: Despite the age variable looking normally distributed, by binning this variable, a set of patterns are found in this continuous variable which are easier to analyse and interpret. The age bins ranges were selected according to (INE, 2019). After this, the variable Age was dropped.
- **AvgRevenueYear** – average of the total revenue per year – allows to compare among the revenue generated by each customer (in **LodgingRevenue** and **OtherRevenue**) considering the time that they have been customers.
- **Personnightsperbook** – average of *Personnights* that the client does by booking – allows to compare the variable *PersonNights* among the clients. After this step, *PersonNights* was dropped.
- **Revenueperroomnight** – average revenue (total spending by the customer) generated by *RoomNight* – allows to understand how valuable is an extra *RoomNight* for each customer. After this step, **RoomNights** was dropped.

- ***BookingsCheckedInprt, BookingsCanceledprt*** *and* ***BookingsNoShowedprt*** - proportions of each Booking component (*BookingsCheckedIn*, *BookingsCanceled* and *BookingsNoShowed*)

The categorical variables **DistributionChannel** and **Age_bin** were transformed to dummy variables to further consider them in our analysis.

3.2.2.2. Import external data source to complement the analysis

The ISO code for countries (ISO-3166-codes, n.d.) was added to the data frame in order to have 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 in a geographical context.

3.2.3. 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. The next step was to plot a correlation matrix to identify and avoid redundancy in the dataset.

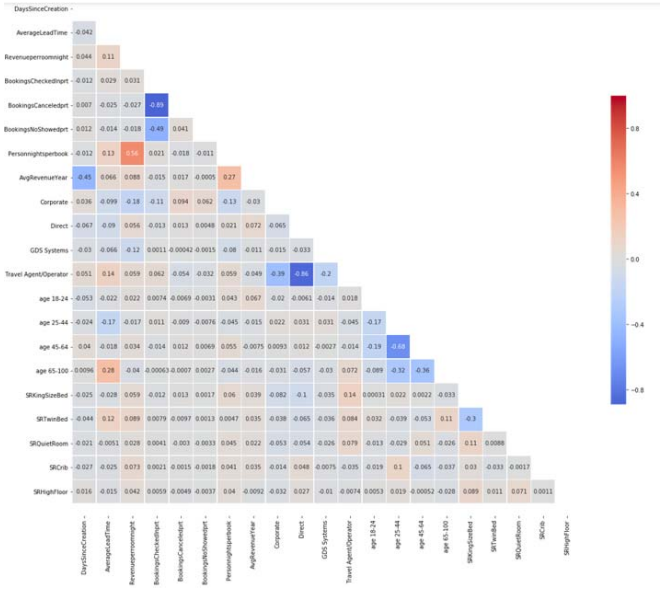

Figure 1. Correlation matrix.

BookingsCheckedInprt and **BookingsNoShowedprt** (-0.49 correlation): 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 and **BookingsCanceledprt** (-0.89 correlation): This high correlation was explained by the low representation of BookingsNoShowedprt. Excepting 31 cases, BookingsCheckedInprt = 100 – BookingsCanceledprt and therefore, BookingsCheckedInprt was kept.

PersonNightsperbook and **Revenueperroomnight** (0.56 correlation): 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.

3.2.4. Data Visualization

After cleaning the dataset, in this section, combinations of different features were plotted to find more insights about the business and the dataset. After the data coherence / discrepancy verification we have, in our database, a total of 75.121 bookings checked in, 219 bookings cancelled and 58 “no showed”.

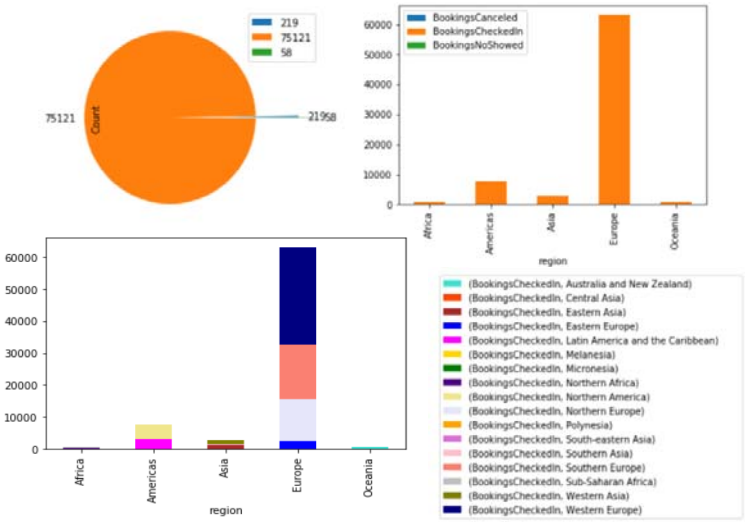

Figure 2. Number of Bookings distributed by region.

The Hotel has a low cancelation a non-show percentage. Most of the checked in customers are coming from Europe, followed by America and Asia. Considering the sub-regions:

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

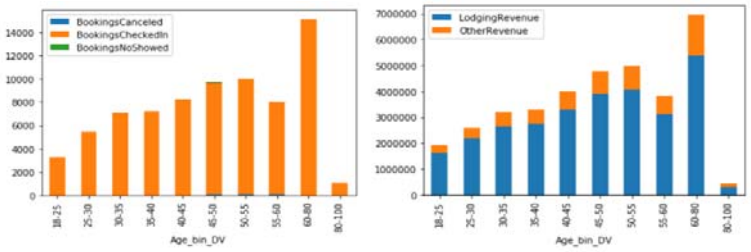

Figure 3. Bookings by age bins.

Looking for the ages, in Figure 3, most of the checked in clients have ages of 60-80. Client with ages of 18-25 have the lower representation in check ins due to the cumulative behavior of customers' bookings. Expenses logically follow the same pattern of bookings, since once they booked, they can spend money. The Lodging Revenue has the bigger representation on total revenue by age bin too.

3.2.5. Outliers detection and treatment

Since K-means is the proposed modelling method for this business case, data points that are far away from the data cloud (outliers) should be excluded to enhance the performance of the clustering technique. Boxplots visualization, which show values outside a defined Interquartile range (IQR) were used for outliers detection. They allow a first glance of the outlier’s distribution. For some features, a second boxplot was created to ensure the right amount of outlies were excluded.

- **BookingsCheckedInprt:** As mentioned before both variables regarding the percentage a person canceled a booking or simply did not show up were dropped, both because of their correlation with this variable and their low representation. Meaning it was expected for most of the values to have

a high percentage. After observing the boxplot, values below 0,6 were considered outliers. This criterion is supported by the idea that customers that do not follow through with more than half of their check ins are not good customers, meaning that is better to exclude from the segmentation.

- **RevenuePerRoomnight:** values over 5 and below 0,75 were excluded. These are customers that either spent significantly less for each night and room they stayed at.
- **AvgRevenueYear:** outliers are very noticeable for this feature. Values above 3.000 were excluded.

After excluding the outliers, the number of records keeped were: 69.447, which is the 62,18 % of the initial records. So, there were 31 outliers detected and excluded.

3.3. MODELING

3.3.1. Principal Component Analysis (PCA)

In order to reduces the attribute set size, but keeping the main variance of the dataset, PCA is performed to create a smaller set of variables. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result (Han et al., 2011). For instance, to achieve the best customer clustering, we use PCA to reduce dimensionality. Before implementing PCA we normalized the Data with the *StandardScaler* module from scikit learn. StandardScaler scales to the unit variance after subtracting the mean of each feature. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

Because the principal components are sorted in decreasing order of significance, the data size can be reduced by eliminating the weaker components. As an output of PCA we receive 15 principal components that cumulatively explain 97.37% of the variance. Having successfully reduced the data to a simplified form with fewer dimensions, the next step is to segment the customers into clusters by performing K-means unsupervised algorithm. Subsequently, following the clustering process we use the Elbow method (Yellowbrick, n.d.) and average silhouette score (Scikit-Learn, n.d.) to find the optimum number of clusters before applying K-means.

The chart displayed on figure 6 summarizes the results of both methods for selecting the optimal number of clusters. Considering both, the Sum of Squared Errors and the average silhouette scores, the number of clusters K = 6 is selected, since the first Elbow-shape for SSE is detected on this point as an increasing on the average silhouette score. After performing K-means, labels for the 6 clusters were assigned, and for instance, a color to distinguish to which cluster belongs each datapoint using the same two principal components for visualizing the scatterplot of figure 7 Were the biggest black dots representing the clusters centers.

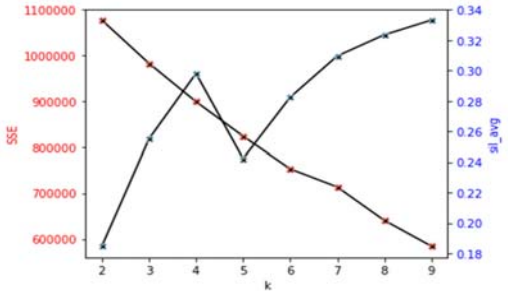

Figure 6. Elbow plot for SSE and average silhouette scores

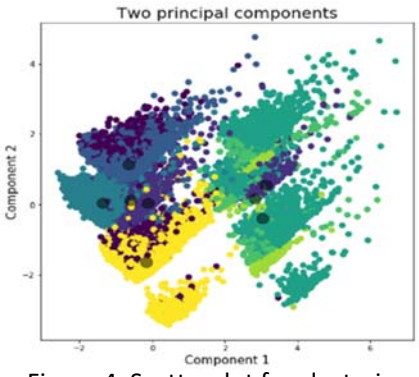

Figure 4. Scatterplot for clustering

4. RESULTS EVALUATION

As exposed in the previous section, the modelling algorithm was performed for 6 clusters, the resulting distribution of customers is the following: Cluster 1 (21.184), Cluster 2 (8.766), Cluster 3 (7.204), Cluster 4 (8.967), Cluster 5 (21.342), Cluster 6 (1.984). The next step was to analyze the loading of each variable over each cluster to identify the customer's profile of each cluster. The results were the following:

Cluster 1: Young customers mostly booking through Travel Agents or Operators
- Almost half of the customer's request King Size Beds and High floors, meaning couples
- Majority of bookings done by Travel Agents or Operators
- Average spending's in hotel
- Cluster with the youngest customers (age 18-24 and age 25-44)
- On average, this group has a smaller lead time
Cluster 2: Adults booking through Direct Channel having the highest spending
- Preferences in King Sized beds and High floors
- Most profitable group
- Almost all bookings done through Direct Channel
- Adults (age 25-44 and age 45-64)
- Smaller average lead time than the average customers
- Most recent customers
Cluster 3: Adults booking through Direct Channel with family doing high spending
- Almost all the people in this cluster request Twin Beds, mostly family
- Almost all bookings done by Direct Channel
- Preferences in Quiet Rooms, Hight floors and Cribs
- Second highest revenue generated
- Adults (age 25-44 and age 45-64)
Cluster 4: Elderlies doing bookings in advanced through TA/O, requesting specific beds
- Almost all bookings done through Travel Agents or Operators
- Elderly (age > 64)
- Customer who do the booking most in advance
- A significant portion of the clients request either Twin or King size beds
Cluster 5: Adults doing booking through TA/O requesting King Size Beds
- Almost half of the customer's request King Size Beds, meaning that has mostly couples.
- Majority of bookings done by Travel Agents or Operators
- Adults (age 45-64)
- Customer who have been with the hotel for a long time
Cluster 6: Adults doing booking through Corporate spending the least on the hotel
- Least profitable customers
- Only cluster whit booking through Corporate
- Adults (age 25-44 and age 45-64)
- Bookings done with the least time in advance
- Customers who have been with the hotel the longest
- In the majority, do not have special request

Table 1. Clusters characteristics

5. DEPLOYMENT AND MAINTENANCE PLANS

This section exposes the strategic plans and marketing recommendations in order to target the already existing customers, now clustered, with high efficiency. However, the marketing plan depends on companies' objectives and managers' decisions. Therefore, the following standard approaches are still flexible to change depending on the company's criterion.

- **Cluster 1:** This cluster is mostly comprised of young and lower than middle aged customers who mostly book with Travel Agents and Operators. Therefore, it would a very promising strategy for the hotel to use online marketing and social media promotions to grow and attract this customer base. Recommended, channels would preferably include famous social media platforms such as Instagram or Yelp that would include number of pictures of King-sized beds and preferences exposed by the model results. Creating and advertising recreation areas for parties or physical activities is also advised.
- **Cluster 2**: Customers belonging to this are should be considered as very important for maintenance perspective, since they provide the highest return on average. Therefore, a delicate Customer Relationship Management (CRM) program could very well target this cluster of customers, staying within a line of direct contact and offering promotions. These customers comprise mostly of people with the special preferences of King-sized beds and Higher floor, therefore, as one strategy they all can be offered a "VIP" bundle that will offer them extra services resulting in higher revenue.
- **Cluster 3**: A very reasonable strategy to increase customer retention, satisfaction and spending for this cluster would be to develop a "family" program. Examples include low cost implementations such as children caretaking programs and playgrounds for families and special romantic room & food services for couples. This cluster is much alike to the 2nd cluster in their special requests. However, they are the highest with crib preference therefore some kind of room allocation program would be appropriate to eliminate the possibility of noise caused by babies in cluster 2.
- **Clusters** 4 and 6 share similarity with regards to their average revenue, homogeneity and channels. A transportation service program could be developed for both. It could provide fast cab services for corporate customers of cluster 6 and shuttle bus services for older touristic customers of cluster 4. Some other utility services also can be implemented like secure provision of WIFI for corporate customers, availability of printing services and increased accessibility for older customers. Lunch services also could be adjusted to be provided faster for Corporate customers and healthier options for elderly customers of Cluster 4.
- **Cluster 5:** A retention program that could further promote customer loyalty would be ideal for this cluster since is the oldest customer group. Approaches include reward and bonus collection for each additional visit or visit hour and post booking contact. Since this highly profitable and loyal cluster, a very efficient strategy could be to try to pivot this clusters preference from TA/O bookings to direct bookings to increase profit margin with minimal effort.

Is suggested to execute this model at least every six months to capture the seasonal changes of the business and update the marketing strategies if needed.

6. CONCLUSIONS

Nowadays, it is important to take advantage of customers data for a better understanding of the business and for not to stay behind against competitors. The proposed customer segmentation with strategies indeed include a Data Driven analysis in order to align the hotels services with the customers' needs based on available data.

Despite of presenting the deployments strategies according to the modelling results, it is recommended to address the clusters' strategies following the Pareto principle to generate the biggest positive impacts investing the least resources.

The most profitable customers are in cluster 2 and 3, but the customers with a highest number of customers are cluster 1 and 5, as the hotel seems to attract more customers with these characteristics.

The cluster 6, is the less profitable and with less customers, therefore investments in the Corporate customers will represent less revenues.

In terms of investment, it is advisable to take first actions over clusters 1 and 5 to expect high returns because of the customers volume. In other hand Cluster 2 can quickly return investments, because it represents customers with the biggest revenue per room per night.

6.1. IMPLICATIONS FOR BUSINESS

Since the marketing strategy focuses company attention on particular target market segments and makes it clear what product characteristics are required for successfully satisfying customer needs, the implementation of a new marketing strategy based on the proposed customer segmentation will have positive effects on the Hotel company:

1) **Short-term**: Have saving by eliminating operations and investments that do not contribute to business growth.
2) **Mid-term**: Improve business profitability because of the accuracy of the offering the correct products and services to the correct customers.
3) **Long-term**: With improved customer satisfaction and innovative products of high quality, the Hotel company can increase the rating on social media and consolidate the Hotel brand.

6.2. CONSIDERATIONS FOR MODEL IMPROVEMENT

During analysis it was noticed that the addition of certain variables like profession, number of children, travel reason, nights spent in total, general satisfaction with the Hotel service. Would improve the results of the segmentation, so we recommend adding short printed surveys at the entrance or digital surveys before accessing the Wi-Fi. Promotions or potential offers would be given to those who complete it as an incentive.

In the coherence verification portion, it was noticed that 177 clients had zero *Revenueperroomnight*. Since this is a transformed variable composed of the total revenue and room nights, total revenue was examined. It was observed that some customers did indeed have zero Total revenues. Meaning they have never spent anything since the first time staying at the hotel. Most likely this represents employees who stay at the hotel cost free. Since they are not customers, they should not be a part of

our segmentation and an additional variable should be created (i.e.: Employee) to identify them, so they can be excluded in further analysis.

In addition, a variable called “Season/Holidays” could be created. It would identify the most profitable seasons for the hotel and could potentially help in the creation of promotions for the seasons with less revenue. This information is vital to increase the understanding of the hotels client’s preferences and what they value the most.

Another possible improvement would be to test other algorithms to the dataset and evaluate their accuracy and performance. Finally, as required for good quality models, outliers were removed from the dataset. After the clusters are created the outliers should be reintroduced by creating a decision tree.

7. REFERENCES

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

INE. (2019). *Estatisticas de turismo 2018*. https://doi.org/10.1017/CBO9781107415324.004

ISO-3166-codes. (n.d.). *lukes/ISO-3166-Countries-with-Regional-Codes: ISO 3166-1 country lists merged with their UN Geoscheme regional codes in ready-to-use JSON, XML, CSV data sets*. Retrieved February 23, 2020, from https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes

Ncr, (, & Clinton, J. (2000). *Step-by-step data mining guide*. DaimlerChrysler.

Rondan-Cataluña, F. J., & Rosa-Diaz, I. M. (2014). Segmenting hotel clients by pricing variables and value for money. *Current Issues in Tourism*, *17*(1), 60–71. https://doi.org/10.1080/13683500.2012.718322

Scikit-Learn. (n.d.). *Selecting the number of clusters with silhouette analysis on KMeans clustering — scikit-learn 0.22.1 documentation*. Retrieved March 3, 2020, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Waida, M. (n.d.). *Hotel Target Market: The Keys to Finding the Best Customer*. Retrieved March 1, 2020, from https://www.socialtables.com/blog/hotel-sales/best-hotel-target-market/

Yellowbrick. (n.d.). *Elbow Method — Yellowbrick v1.1 documentation*. Retrieved March 3, 2020, from https://www.scikit-yb.org/en/latest/api/cluster/elbow.html