HOCHSCHULE FURTWANGEN UNIVERSITY

**Summer Term 2018**

# Explorative Analysis of Transaction Data for the Tourism Industry

## Thesis

Submitted in partial fulfillment for the degree of

## Master of Science

in Business Consulting

in the Faculty of Business Information Systems

Hochschule Furtwangen University

by

**Ivan David Rojas Gomez**

---

| | |
|---|---|
| Reviewers: | Professor Dr. Pavel Rawe<br>Professor Dr. Holger Ziekow |
| Submission date: | 8th October 2018 |

# Affirmation

I, Ivan Rojas, hereby declare in lieu of an oath, that I personally composed the following thesis independently and with no assistance from a third party.

All sources are fully cited.

Furtwangen, 8th October, 2018

_____

# ABSTRACT

This study examines the association among several, internal and external attributes derived from the practice of revenue management that can influence room pricing within the hotel industry. Previous studies have immersed in the use of sophisticated algorithms to improve forecasting of demand, segmentation of customers, prediction of cancellations, advance bookings and so forth but the idea of this approach on top of finding significant correlations, is to build on that information by incorporating deep learning tools as a feasible technique to ease the selection of prices ranges as a reference for expected seasons. Transaction data has been obtained from a hotel chain formed by five different properties located in United States of America. All datasets are freely available on internet and are found in CSV format. Files have been initially edited in Microsoft Excel and then fitted into RStudio the open source software used for the present analysis. The results in the correlations phase indicate a dependency among the pairs of variables analyzed and that the more predominantly rate is between 80 and 150 dollars among all hotel size properties. It can also be pointed out that CRO/Hotel was the preferred distribution channel. Moreover, the study refutes the impression that bookings purchased with anticipation are always more likely of having a cheaper rate than others booked on short term notice. With the current set of numerical variables a supervised predictive price model was obtained and its outcome evidenced more accuracy than conventional linear regression techniques. Finally, a classification model displayed a selection of factors based on the relevance when explaining room prices as well as the probability for its occurrence.

# TABLE OF CONTENTS

# TABLE OF FIGURES

# INDEX OF TABLES

# LIST OF SYMBOLS AND ABBREVIATIONS

Hospitality Industry terminology

| | |
|---|---|
| CRM | Customer Relationship Management |
| CRO | Central Reservation Office |
| CRS | Central Reservation System |
| ERP | Enterprise Resource Planning |
| GDS | Global Distribution System |
| MAE | Mean Absolute Error |
| ML | Machine Learning |
| PMS | Property Management System |
| POS | Hotel Point of Sale |
| RM | Revenue Management |
| VIP | Very Important Person |

# 1. INTRODUCTION

## 1.1 Motivation

Over the last four decades the hospitality industry has been subject of diverse cost control and revenue optimization approaches to optimize resources while increasing profits. One of those approaches had its roots in the aviation industry in the late 1970's. The successful strategy that led to a significant improvement in profits of a major air carrier in EE.UU, was to accurately forecast empty seats, selling in advance its full fare tickets and in addition, offering the remaining spare seats at lower prices than its main competitors. This method of pricing known as a Revenue Management – RM or yield management became an essential instrument adopted by many other industries in the subsequent years.

A particular characteristic of RM is that it categorizes customers in different segments and then a customized product or service with a specific price is allocated to them. From this context, the term "yield" is derived and basically means discourage customers, who might not be appealing to the business -from the revenue point of view-, by imposing some restrictions such as: rates, room types or distribution channels.[1]

With the advent of the internet in the early 1990's, rates and products among different providers were more easily traced and the restrictions set to filter customers became insufficient. Up to this point researches on the field of RM had focused on developing practices involving the design of products, forecasting demand and classifying target clients. From this stage, a more sophisticated practice, capable of dealing with the transparency of prices emerged from the use of internet,

[1] SNAPSHOT TEAM, "Understanding the Basics of Hotel Revenue Management," Snapshot Travel, November 16, 2015, https://blog.snapshot.travel/understanding-the-basics-of-revenue-management (accessed June 17, 2018).

took place. It incorporated the customer's behavior, characteristics of the products and the market environment.[2]

Although the principles of RM are cross wide utilized by many type of business through diverse economic sectors such as; airlines, car rentals companies, financial services, media and telecommunications, cruise ship lines, theatres; amongst others, the focus of this dissertation will be on the hospitality industry and more precisely hotels as data sets containing historical transactions from a major hotelier are available to conduct a practical analysis. Furthermore, RM is well established in this industry and, after airlines, it is considered an early implementer of the principles outlined.

As RM practice has made possible to influence the demand by effectively interpreting the customer behavior upon a complex definition of factors such as; products diversification, customer segmentation and distribution channels, the focus of this dissertation will be the analysis of their impact on pricing policies within a real business case. Those correlated factors, provided by the RM system, could eventually become the input of a predictive technique that aims to produce a valid model for pricing as well as an effective technique to classified attributes according to their relevance in price formation.

## 1.2 Research Objectives

This study has the following four main research objectives:

Objective 1. Analyze related attributes by selecting informative pairs of categorical variables that will be contrasted graphically and statistically to assess their relevance in pricing formation.

[2] Tudor Dan Bodea, "Choice-Based Revenue Management: A Hotel Perspective" PhD diss., Georgia Institute of Technology, 2008, https://smartech.gatech.edu/bitstream/handle/1853/24739/bodea_tudor_d_200808_phd.pdf

Objective 2. Perform a modern deep learning technique by fitting a model with representative independent attributes correlated with the target of interest, to identify the foundation of an algorithm capable of both; finding patterns based on historical data and providing a predictive value for the dependent variable, price, with low error.

Objective 3. Explore the influence of external factors of the business, through their characterization and integration into a predictive technique to detect possible hierarchical levels of importance amongst variables and the way they impact price intervals.

Objective 4. Propose a complementary method for evaluation of factors that impact the business' profitability by implementing a quantitative method that recycles data generated by the revenue management practice and provides a systematic price reference.

## 1.3 Thesis Structure

This dissertation is divided in eight chapters. The second chapter is dedicated to the definition of revenue management and the special characteristics of the hotel business in the light of current technical literature about these topics. In particular key elements linked to the RM system are studied and its implementation process within the hotel. Then a standard classification of type of hotels, customers and reservation channels is followed. The third chapter comprises an approach to the research by explaining the concepts that support the methodology used in the experimentation phase. The three main models are correlations, neural networks and decision tree. Chapter four explains the procedures conducted to collect data and adjust it for fitting the predicting models. It also makes a description of all

attributes which are summarized according to their type in a table format. In the chapter five, data is analyzed by fitting the proposed models. Functions and coding utilized in RStudio are shown together with the specific set of variables incorporated. Results come in the chapter six; they correspond to the outcome of the experimentation step reported by RStudio and are exhibit through diagrams and tabulation to facilitate its analysis. Discussion is found in chapter seven where main findings are explained in detail and chapter eight contains the conclusions.

# 2. REVENUE MANAGEMENT IN THE HOTEL INDUSTRY

## 2.1 Review of the definition

From a practical perspective Revenue Management - RM can be defined as "*The wide range of techniques, decisions, methods, processes, and technologies involved in demand management"* (van Ryzin and Talluri, Pag. 2, The Theory and Practice of Revenue Management, 2005).

This concept has to do with the fact that firms can influence the demand coming from their clients in order to maximize its revenues. To do so, customers are segmented to carefully analyze their willingness to pay for a customized product which is available in restricted amounts (i.e. rooms in a hotel, seats on an aircraft). Additionally, seasonal factors are taking into account as products are treated as perishable goods, susceptible of being labeled with many different tag prices during its validity. The logic behind is that businesses are committed to sell up to the last unit of its inventory at the highest possible price (conditioned by competition and cost structure). But when its maturity gets closer prices should be reduced.

In essence, demand management in hotels refers to the decisions regarding; what customers to serve, when and what product to offer, price allocation, distribution channel and marketing needed to increase revenues under uncertain conditions.

Demand management and its transcendence in the industry has been made possible thanks to the advance in science and technology that enable humans to automate operations that manually would take years given its enormous complexity. RM is linked to multiple disciplines including; economics, statistics and operations research, fields with important scientific advances, which in turn has led to industry practitioners, to develop more accurate forecasting and calculate optimal solutions based on demand models incorporating into the analysis real market conditions.

On the other hand, RM finds support in information technology and related devises (modern data bases, processors, high speed networks, software, etc.) as they facilitate data collection and its storage, calculation of multiple transactions with huge volume of data by using sophisticated algorithms and, lastly, helping decisions makers with the reporting and implementation of high quality decisions.

## 2.2 Core Components of Revenue Management

A summary of the main variables in Revenue Management can be seen in figure 1.[3]

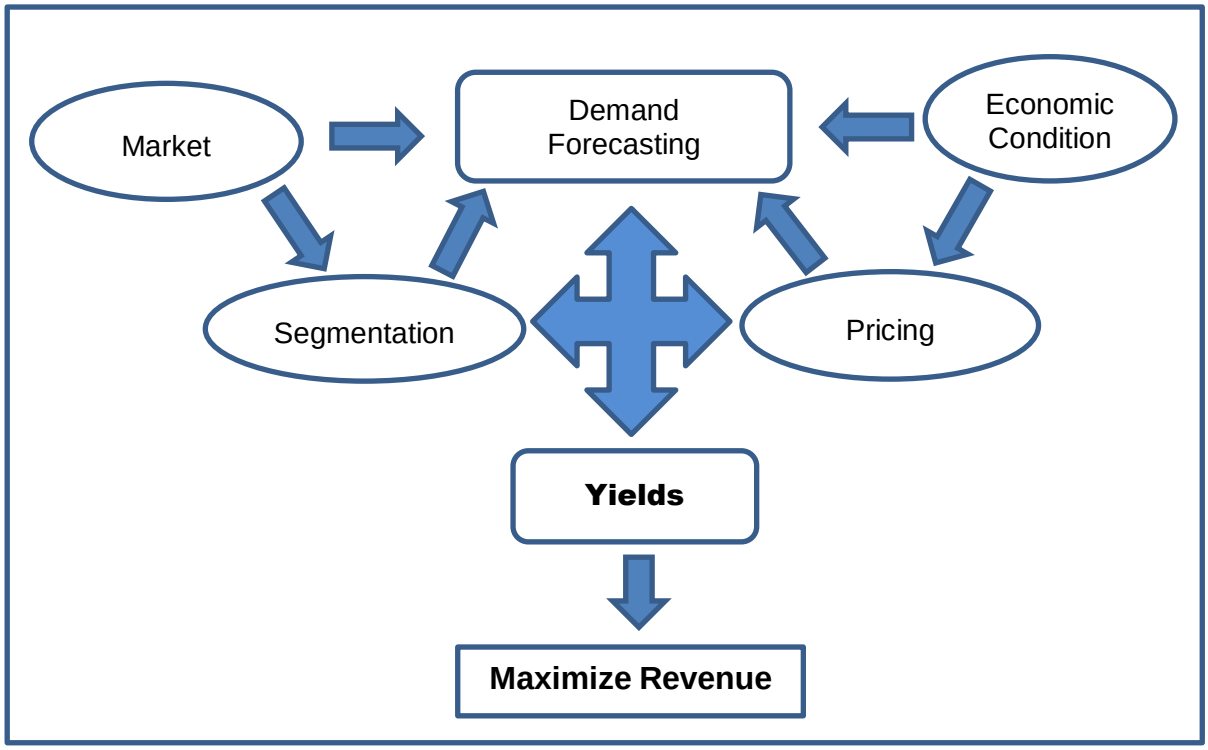



**Figure 1. Core components of Revenue Management**

Economic conditions are influencing market and pricing which in turn affect the way customers prioritize purchase decisions and set their expectations. Demand is

[3] Debra Adams et al., *Revenue Management: HOSPA Practitioner Series* (Bournemouth: Wentworth Jones Limited, 2013).

constantly updated and its forecast plays a key role determining inventory levels or yields. With the aim of maximizing revenue, a huge deal of calculations must be carried out to improve the performance of the above components and this can be partially achieved with the use of machine learning algorithms. Machine learning - ML is an extension of statistics and computational analysis that enable computers to learn as a matter of finding statistical symmetries or patterns in data.[4] At this point a brief description of each component of RM is made:

- Segmentation. Divide guest based on diverse aspects such as purpose of travel, age, gender, prefer payment method, geographical precedence, income level, marital status, new or returning guest, etc. ML also known as deep learning tools allows the application of clustering and classification to offer customers tailored products to suit their needs and budget.
- Demand. Multiple methods have been used to forecast demand for hotel services including linear regression and moving average up to sophisticated algorithms which can be fitted with information of competitor's historical prices, public events, room types and so forth.
- Yields. This term refers to inventory control and the way it needs to be executed to optimize profit and increasing revenues. The general goal is to fully capitalize on revenue opportunities during peak demands periods and maximize occupancy in low demand seasons and this is achieved when fixed costs are covered by managing the inventory. This is particularly important when selling rooms as the proportion of fixed costs is high compared to the proportion of variable costs. The latter ones only account a small fraction of the selling price and the remainder amount is allocated to pay off fixed costs and contributing to the profit.

[4] Taiwo Ayodele, *Types of Machine Learning Algorithms*, https://www.researchgate.net/publication/221907660/download, (accessed August 30, 2018).

- Pricing. It is characterized by flexible room rates explained by a price strategy which takes into account customer segmentation, occupancy rates, supply and demand analysis.

This dissertation will make emphasis in pricing identification by using ML tools as different attributes collected for the analysis phase are comprising data from RM systems.

## 2.3 Ideal conditions for revenue management

Airlines, hotels, car rentals or ship cruises are examples of businesses which exhibit unique characteristics that make RM implementation a successful practice.

- Fixed supply. Hotels have a limited amount of rooms available and increasing its number requires a significant length of time planning, designing and constructing
- High fixed and low variable costs. Fixed costs are depicted by salaries, taxes, or debts and variable costs are associated with selling one more room (e.g electricity, water bills)
- Interchangeability of products. It refers to the essential product offered by hotels. Rooms. Value added by providing better services, marketing or amenities can represent additional revenue but also customers might find a similar deal with the competitor
- Segmented markets with differing price sensitivity. Guest segmentation enables hotels to classify their customers according to the purpose of their stay and willingness to pay different rates. Higher rates and short stays are commonly associated to business travelers whereas flexible dates and discount rates are linked to leisure guests

- Seasonality of demand. Hoteliers are capable of forecast demand based on historical patterns and can also make projections according to the season.

## 2.4 Revenue Management System

It is the system that controls price and capacity simultaneously. Its generic process can be divided in four different steps:[5]

1. Data collection: Represented by historical data. It can be numerical or textual.
2. Estimation and forecasting. For estimation of demand based on parameters defined by the firm.
3. Optimization: It is a set of rules to determine prices, discounts and basket of products that optimize revenue. It must be performed periodically.
4. Control: It is the method used by the company to monitor sales which in most cases are done thorough different distribution channels.

The sequence of the system is the same but it can be carried out at different intervals on time according to the business requirements, the speed and volume of data collected. Figure 2, shows the process flow in the RM system.

Starting from the top, data collection layer receives records and figures coming from the customer historical databases as well information regarding the products and its prices. This data is processed by the forecaster (Estimation/Forecasting layer) and then passes into the optimizer (optimization layer) which determines the best combination of products and prices to maximize the revenue. Finally this information, feeds the reservation system which control sales through different distribution channels.

[5] Garret van Ryzen and Kalyan Talluri, *The Theory and practice of Revenue Management* (New York: Springer, 2004), 17.

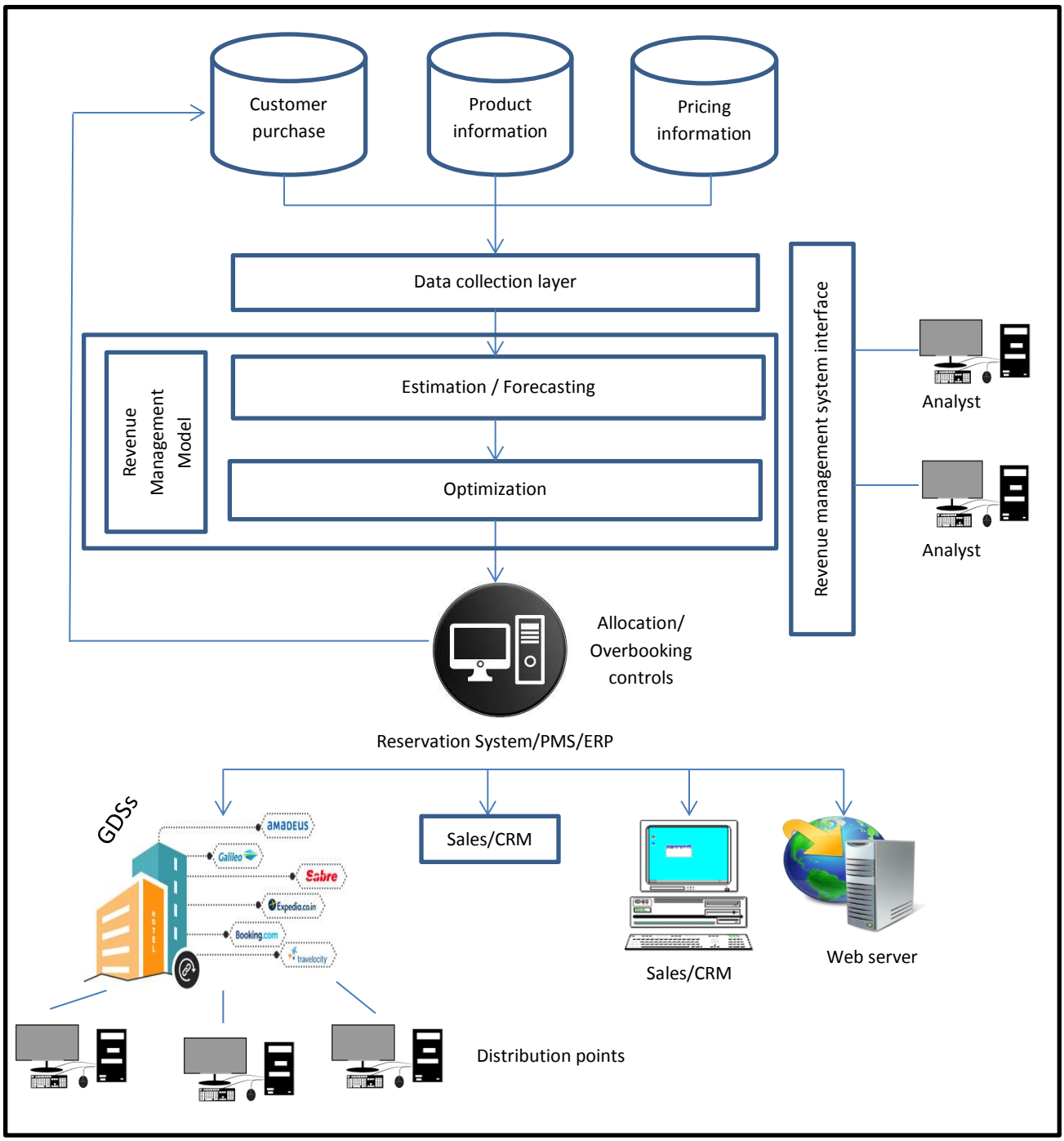



**Figure 2. Revenue Management System (adapted from Ryzen and Talluri, 2004)**

## 2.5 Customers, products and Rates

Hotels can be categorized as a resorts, business, leisure or combination of them. They might be distinguished by size, location or even ownership.

Each customer class has a different reaction to variation in prices and its segmentation is done following the criteria of each hotel brand but, in general lines, they can be partitioned as individuals and groups where the purpose of their stay varies. The first ones are commonly classified as independent travelers who booked their own room for business (corporate class) or leisure. A second segment can be groups shaped, for example, by tours or staff from a company that the hotel has an agreement with (corporate and tour) buyers.

Corporate guest tend to be less sensitive to prices. This type of customer stays at the hotel because need arises due to their work requirements and are not traveling lured by a reduction in room price. Additionally, these travelers are willing to pay higher rates due to their usual sudden trips that shorten the length of time between the booking and the checking dates, reducing their chances to obtain discounted rates. These travelers are also keen to pay more for location and comfort.

Unlike Corporate guests, leisure customers typically plan their trips well in advance, shop for the best rates and are more flexible in terms of destinations and dates. Even should prices not suit their budget, they might not decide to travel. Leisure facilities must strive for become more appealing to them by offering heavy discounts in low demand or off-season periods.

Group guests depict characteristics of the other two categories. From the leisure side they are coming for instance, from associations, fraternities or clubs. Corporate groups can belong to government entities, companies or unions. Their sensitivity to price variation also varies according to the classes above explained.

Some products offered by hotels, which address specific segments of travelers are: corporate (aimed to staff from companies and planned for long term stays), vacations, weekend packages and walk in clients among many others and its price

differentiation will depend on multiple factors including; seasonality, channel of distribution and advance-purchase.

From the point of view of revenues for hotels, rooms are considered the main source of income despite the fact that other sources also contribute with significant sales including; food and beverages, functions and events, leasing of venues, spa and gambling. However, RM systems often work only with the revenue derived from rooms selling.[6]

As mentioned above, rooms are significant contributors in terms of revenue and raw material for the RM systems hence its classification into numerous types as well as its corresponding rates. Some examples of room types, which are actually listed in the data set of this study, are suite and standard rooms which in turn are classified in smoking and nonsmoking, with double or king beds size. Its price falls in different categories referred in the data set as a rate codes with its corresponding description.

Rate codes linked with rack rates are those publicly advertised with the highest price for a particular room type. Discount rates will descend as a fraction of the rack rate and a customer can be entitled to those special rates upon his membership status (i.e. VIP), advance purchase or negotiated discounts due to his connection with large organizations (i.e. government, company, clubs etc.). On top of this classification, travel agencies can negotiate wholesale rates which are normally lower than the corporate type.

## 2.6 Elements of Revenue Management Practice in the Hotel

Four different common elements for the practice of RM in the hotel industry can be identified.[7]

[6] Garret van Ryzen and Kalyan Talluri, *The Theory and practice of Revenue Management*, 525.

[7] Ibid., 526

### 2.6.1 Booking Process

Bookings are made normally with the hotel or through Global Distributed System – GDSs. Their lead time is typically 18 months or less yet most of the reservations are placed just few days before the check in date. In order to secure a room customers are required to disclose their credit card details, paying at front full or partial rate or even are given the opportunity of cancelling free of charge before their arrival. These practices vary considerably according to the country or region.

When customers request a rate they are normally offered either the cheapest rate available or a much higher with the possibility of bargaining. The latter practice is in disuse due to discomfort caused by the negotiation process and the existence of internet portals where customers are able to choose by themselves, from a comprehensive list, the package and price that suit their needs.

Corporative clients are usually offered, by the hotel agent, more expensive rooms/packages if the initially selected option is sold out.

### 2.6.2 Property Management System PMS

It is an administration system that controls reservations, availability and occupancy of rooms. PMS also records many transactions in real time including sales of meals, beverages, lease of venues and functions, inventory and billing. It is also linked to other applications from external GDSs or hotel point of sale (POS) which are terminals used to process payments.

PMS systems are essential tool for reporting and communication. They enable the collection of vast amount of information which is then displayed in a consolidated and customized manner and its connectivity capabilities with GDSs allow hotels track reservations done by third parties.

### 2.6.3 Overbooking, No-Show, Cancellations

The hotel is overbooked when a customer with a valid reservation arrives on the check in date and no rooms are available to host him. Under those circumstances the hotel must walk the client to a different location. In general, this is a rare event as most hotels tend to be conservative when allocating rooms nevertheless establishing limits for overbooking is a common practice.

Ratios of overbooking are normally set depending on seasons or dates where special events, around the location are taking place, what in turn helps to determine the number of guests expected. Particularly, leisure hotels are prone to display more rooms than they actual capacity in an attempt to attract more clients from tours operators, hoping that not all vacancies offered will be filled up.[8]

No-shows are customers who neither use nor cancel the booking. And those who call the hotel or travel agency to give notice for their not attendance before the check in date, are known as a cancellations.

### 2.6.4 Capacity Controls

Capacity control, often associated to the length of stay, is in place to help hotels to retain or to walk away customers with different spending power. It can be determined by a *minimum length of stay,* which occurs when the hotelier sells only long term stays or multiple nights in order to secure the income from a room especially in periods of high demand, discouraging demand from customers who intend stay one or just few nights.

The opposite practice is called *maximum length of stay*, and by applying it, hotels favor customers for short term stays with high rates per room displacing those long term stays with low to moderate rates room per night.

[8] Ibid., 530

*Close-to-arrival* controls enable hotels restrict the stay to a specific date. Finally, a control called *open-for-day*, which is not widely used, will limit the stay only for a day use.

An automated revenue management system, fed with the above controls, will make forecasting and optimization of availability and revenue multiple times a week and the procedure tends to dramatically increase as the reservation usage nears. This happens because most bookings occur during the last days previous to the check in date.

## 2.7 Break-Even Point

It occurs when the business has neither profit nor loss. In this scenario revenues equal to costs. Hotels in particular have high fix costs including accounts payables, licenses, permits, taxes and fix salaries[9]. Thus, high level of occupancy is not required as long as fixed costs are reduced. Likewise, diversifying the source of income by selling food, beverages, functions, spa and additional products alike, will directly remove pressure on room sales to cover all the hotel expenses.

At Break-even point, no profit is achieved and all revenues are used to pay off debts, fix and variable costs. Once the income exceeds expenditures, a profit starts building upon any additional sale.

Break-point points are stated in occupancy percentages and its value can further be reduced by applying customers' segmentation and recurrent variation in room rates, common practice in revenue management.

[9] Gary K. Vallen and Jerom J. Vallen, *Check In Check Out: Managing Hotel Operations* (Boston: Pearson, 2018), 7.

## 2.8 Hotel Revenue Management Process

With some variations among different authors, the process can be comprised in seven stages as illustrated in figure 3.[10]

- Goals. These are set by the mangers of each cost center (food and beverages, functions, rooms, spa, casino, etc.) and then communicated to the general administration who will be in charge of prioritize them according to short, medium or long term attainment.
- Information. Collection of operational data, source by the hotel's marketing system containing customer's segmentation and historical records of products, prices, campaigns, etc.
- Analysis. Operational data and indictors are analyzed by management to identify potential trends as well as demand for services for the coming days or weeks.
- Forecasting. Estimations and forecasts about demand, supply and other metrics are developed and become precondition for RM decisions.
- Decision. A vast amount of calculations take place in this stage upon the forecasts and controls loaded from the previous step. Based on the recommendations provided by the algorithms of the RM system, managers would take actions such as modifying levels of prices, limiting availability of a certain type of rooms, restructuring packages or implementing capacity controls. In theory, the decision process can technically be automated by the RM system but in practical scenarios, managers, continue making the final judgment.
- Implementation. In this stage the decision is communicated to the hotel staff particularly those in the reservation department. In order to implement

[10] Ivanov Stanislav, *Hotel Revenue Management: From Theory to Practice*, (Varna: Zangador Ltd, 2014), 34

it, the staff requires training to acknowledge a new product or strategy in addition to sales techniques. The hotel web site and GDSs will also require being updated with terms and conditions associated to the latest announced package or new guidelines.

- Monitoring. In this final stage, the RM process is evaluated in all its stages in order to introduce improvements or capitalize opportunities. It also involves following the implementation of decisions or plans and putting in place corrections when needed. Furthermore, the review of the whole process is required even when main goals were achieved as they might have been effortlessly accomplished.

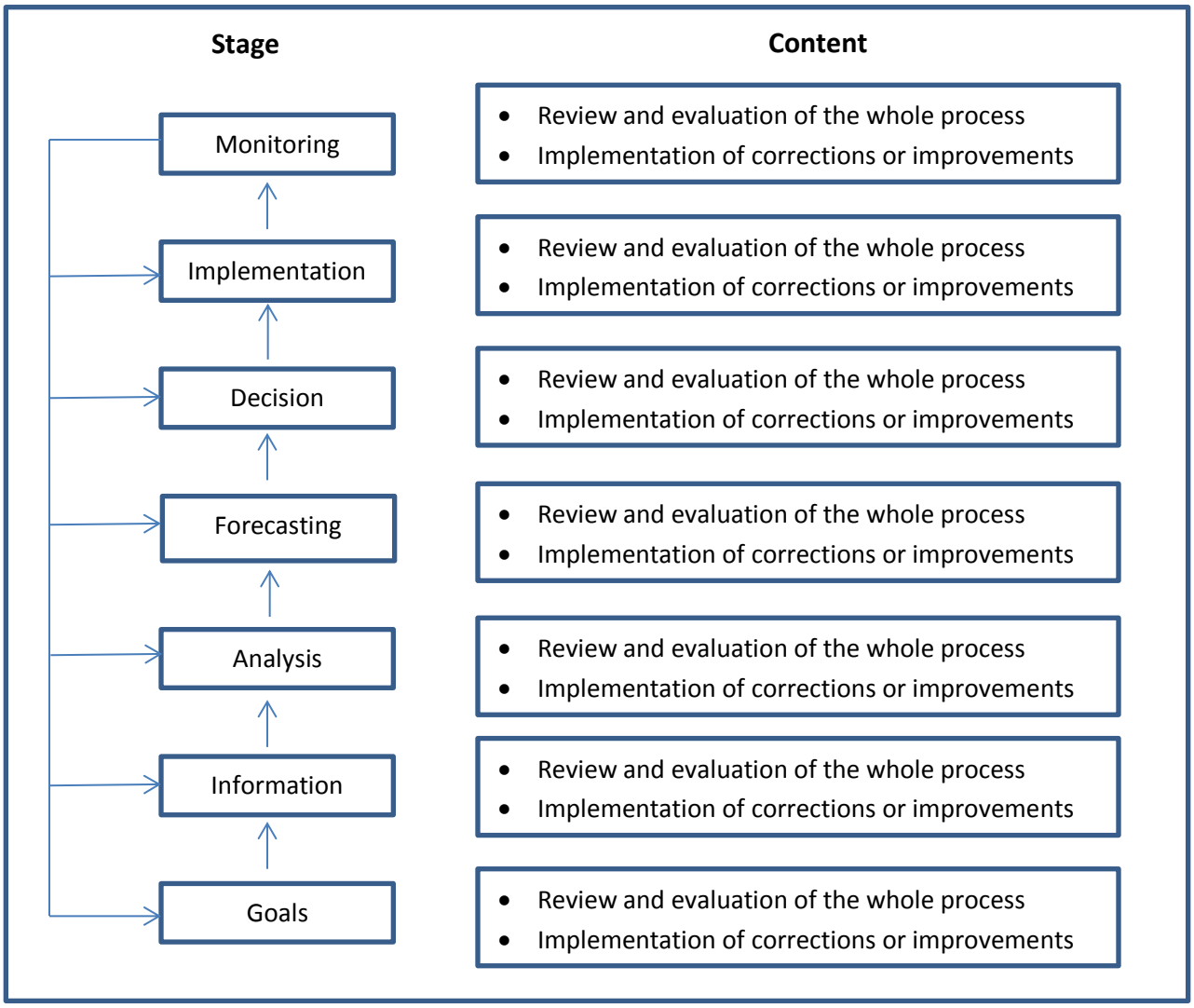



**Figure 3. Revenue Management Process (adapted from Ivanov, 2014)**

## 2.9 Special Characteristics of the Hotel Business

### 2.9.1 Perishability

Unsold hotel rooms are assets which do not bring benefits despite its usability. Therefore, are unable to generate income to cover for its own maintenance or contribute with profits to the organization. An unsold room - night represent for the hotel a permanent missed opportunity to become more lucrative so that the quest, among managers, to have a fully booked lodging every day and season.

### 2.9.2 Location

It is the geographical place or neighborhood where the property is situated. As a fix location, it is an important criterion for the hotel success and its analysis requires the intervention of sales and marketing.

### 2.9.3 Fixed Supply

The total availability of rooms is fixed. Unlike other industries, the amount of rooms cannot increase or decrease in quantity according to demand.

### 2.9.4 High Operating Costs

Hotels require large capital investment and are labor intensive. Costs generated will remain high despite the occupancy levels.

### 2.9.5 Seasonality

Depending on the type of hotel and the market served, cyclical factors affect the hotel. Leisure or business class must face the challenge of keep high levels of occupancy through the week and year. In the first case, hotels are expected to have massive flow of customers on weekends whereas for those hotels orientated to business travelers, the weekend can depict a substantial drop in sales.

## 2.10 Traditional Classifications

Hotels can be classified by: Size, class, type, plan and variation of themes. In the following paragraphs each category will be briefly explained.[11]

### 2.10.1 Size

It is measured by counting the number of rooms for sale. It is important to note that the total of hotel rooms is not a good indicator of the size. Particularly older hotels tend to use some rooms as offices or for storage. Other hotels advertise more rooms than they have actually available.

### 2.10.2 Class

This category comprises two criteria; average daily rate and rating systems. The first one is associated with the magnitude of the rate linked to factors such as; comfort, elegance, room service, valet parking, sauna, etc. It is expected, although not always true, that the higher the rate the better the hotel.

[11] Ibid., 8.

Rating systems is the average score achieved by the hotel granted by its customers. They are commonly published and freely accessible to anyone. Most of those scores are standardized within the same country but they might lack of validity overseas. In some cases, the score is assigned to a multiple factors including; cleanness, quality service, rooms, food, location, price, facilities, etc.

### 2.10.3 Type

Usually hotels type fall in one out of three subdivisions; commercial, resorts or residential. Commercial hotels predominantly, offer short term accommodation for transient (temporal) visitors. Their services are addressed mainly to business people or serving as a conventions facility. Normally are located in urban areas or near to markets, business centers and transportation facilities. Because of its nature, these properties present low levels of occupancy on weekends.

Resorts normally host social guests who are in search of leisure activities. Additional amenities make part of their portfolio including conference centers, spas, pools, guided tours, tennis clubs and golf spaces to particularly attract groups of travelers. They experience higher number of visitors during weekends and holiday periods.

Residential hotels have a permanent occupancy given an agreement that creates a special relationship land lord – tenant. In some cases a commercial hotel can allocate a certain amount of rooms or a section of the property to permanent guests.

### 2.10.4 Plan

A hotel can be classified by plan when is possible to know which meals are included, if any, in a specific room rate. The rate will be higher if it combines food and room. Some examples are *continental breakfast, afternoon tea*, *English breakfast*, *bed and*

*board* and *all inclusive*. The latter one takes place especially in resorts when drinks, food, room, tips and activities are included in one price.

### 2.10.5 Variation of Themes

Some hotel owners equip with unique features their properties by offering an historical or thematic experience to their visitors. Others implement innovative business models to add up services to the existing portfolio of lodging and customer segments. Examples of variation themes are Bed and Breakfast (B&B), Airbnb, boutique and trophy hotels.

## 2.11 Reservation Channels

Traditional distribution channels such as travel agencies and central reservation offices are being displaced by online reservation systems and mobile applications and this trend is supposed to continue.[12]

### 2.11.1 Travel Agent

As mentioned above their roles are decreasing in importance but still are sourcing hotel with bookings. Travel agents work on commission basis and a contract with the hoteliers.

### 2.11.2 Central Reservation Center

Normally is associated with the Central Reservation System (CRS) and the Central Reservation Office (CRO). Both terms are interchangeable nowadays but technically

[12] Vallen, *Check In Check Out,* 134.

the CRS is an electronic system connected to the hotel's availability interface and website as well as the Global Distribution Channel (GDS). The CRO is referred to a physical office where members of the hotel's staff interact with customers.

### 2.11.3 In-House Reservations

This is the most widespread distribution channel used among hotels and it takes place when the customer contacts directly the hotel reservationist for general queries regarding accommodation and rates. It is important to point out that the information received by the client in this approach can differ from the obtained by using a Central Reservation System (CRS) as the staff in-house is up-to-date and better informed about the property.

### 2.11.4 Internet and web-Based Reservations

Internet is one leading tools when customers want to browse, compare and secure a room due to its versatility by customizing the searching criteria and the possibility of executing transactions over the network in real time.

Since internet is expanding as a reservation instrument, so does online marketing hence the importance of hotel websites. Through investing in search engine optimization, interactive maps and updated website contents rich in images and comprehensive information, hoteliers have developed interfaces that improve the user experience which in return can generate substantial profits.

# 3. CONCEPTUAL MODEL AND METHODOLOGY

## 3.1 Introduction

Three different conceptual methodologies that support the idea of price identification along with RM theory will be implemented to analyze the available data sets. The first one is introduced in Section 3.2. This technic measures the strength and direction between variables. Neural networks, the second technic, described in Section 3.3, is used to infer significance and finding patterns with numerical data and it works well with non-linearity conditions. The third method, called decision tree, is presented in Section 3.3 and will handle both numerical variables and factors (text variables) to determine possible interactions and price definition.

In a general context, the latter two techniques are dealing with a supervised classification problem. This is because of the presence of a target attribute, as well as training data used as baseline to be compared against the predictive values reported by the algorithm.[13]

## 3.2 Conceptual Model – Correlation Coefficient

It is a coefficient that measures the association between two variables and rages between -1 and 1. If both examined variables have a linear relationship, the coefficient will be 1. They can also have an inverse relation depicted by -1 or not linear relation at all represented by 0. The Pearson coefficient (product – moment)

[13] Foster Provost and Tom Fawcett, *Data Science for Business* (Sebastopol CA: O'Reilly, 2013), 45.

was developed by Pearson in 1986 and the name of correlation was introduced by the first time in 1988 by Galton.[14]

The Pearson coefficient is widely used and is illustrated with the following formula:

$$r_P = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}, \qquad (1)$$

$\bar{X}$ and $\bar{Y}$ are the sample mean of both variables *Xi* and *Yi* that can be contained in a data set.

Numerical variables will be simultaneously evaluated by using this conceptual model. As a result, a matrix enclosing all Pearson coefficients will be generated to measure their association strength and direction.

A different approach for examining relationships is applied for categorical variables. In this case, factors are paired and displayed in a matrix called contingency table that shows its frequency distribution. These results are then presented in a bar chart.

Finally, a Chi-Squared test with significance level set to 0.05 is conducted to determine whether the two assessed variables are independent or non-independent. A Chi-Square is a hypothesis test which determines the existence of effect on a set of two categorical variables. The association occurs when there is a level of significance. Otherwise differences are explained by chance[15].

RStudio does all the calculations once the contingency table is inserted and in general the process follows these steps:

[14] A.G. Asuero, A. Sayabo and A. Gonzalez, *The Correlation Coefficient: An Overview* (Taylor and Francis Group , LLC, 2006).
[15] David Stockburger, "Introductory Statistics: Concepts, Models and Applications," http://davidmlane.com/hyperstat/chi_square.html (accessed August 14, 2018).

1. It must be stated a null (Ho) and alternative (Ha) hypothesis
2. A level of significance is selected (usually between 0.01 and 0.05)
3. Decide whether categorical variables are or not dependent

Example:

If Ho = the two variables are independent

Ha = the two variables are dependent

Level of significance, $\alpha = 0.05$

Degrees of freedom are obtained through the formula:

$$df = (r - 1).(c - 1) \qquad (2)$$

r and c represent the total number of rows and columns of the dataset, respectively. For this example df equals 2.

The generated test statistic represented as X-Squared is as follows:

X-Squared = 5.8554993, df = 2, p-value = 0.05351734

As p-value is greater than the level of significance α, then the null hypothesis cannot be rejected, therefore, is accepted that these two variables are independent from each other.

## 3.3 Conceptual Model – Neural Networks

Neural network is a machine - software that process information. It has been inspired by the human brain and its interconnected neurons which in turn allow humans to perform certain calculations with a superior degree of accuracy and speed. Some of these calculations (pattern recognition, perception and motor control) are executed in parallel and in a nonlinear way. Furthermore, the brain in humans can provide representations of their environment and consequently people can adapt their behavior according to changing conditions. This ability for learning is simulated by neural networks that it can also model the system in which human brain executes a specific assignment.[16]

The procedure for learning followed by neural networks is called a learning algorithm. This function modifies the synaptic weights of the network to accomplish an objective.

### 3.3.1 General characteristics of Neural Networks

- Nonlinearity. The neurons, which make up the neural network, have a nonlinear interconnection between them and along the network.
- Input-Output Mapping. The neural network is uploaded with a set of examples or training examples. As a result, its synaptic weights are modified, a process called *supervise learning*. Each example is a distinct signal with an identified objective or output. An example can be an email spam filter. The input or example data are words in the body of the document associated normally with spam and the output is the classification of whether the email is spam or not. After many examples have being processed by the neural network, it can eventually recognize what content makes it likely to be a

[16] Simon Haykin, *Neural Networks and Learning Machines* (Hamilton: Pearson Prentice Hall, 2009).

spam.[17]

- Adaptivity. It means that a trained neural network operating in a specific task can be adjusted to work under different conditions or tasks.
- Uniformity of Analysis and Design. Neural network notation is the same in all contexts. It has universal acceptance as neurons have a common representation making possible sharing theories, data analysis and learning algorithms.

### 3.3.2 Neural Network Structure

- The activation function. This function controls the output according to the size of the input. If the latter one is greater than a certain value then the condition of the output changes to a number between 0 and 1. This activation function is commonly depicted by the sigmoid function:

$$f(z) = \frac{1}{1 + exp(-z)} \qquad (3)$$

- Nodes. Networks are connected with layers of nodes as observed in the below diagram:

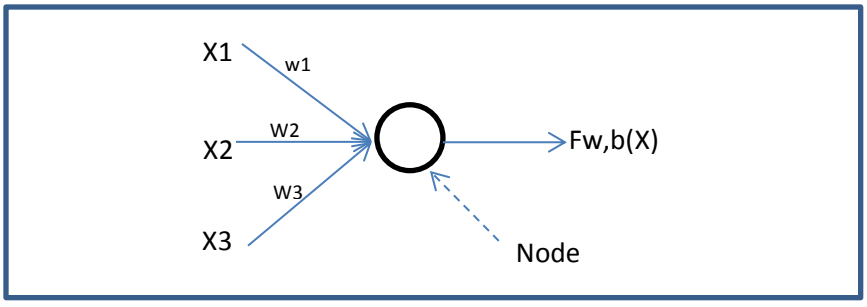



**Figure 4. Node and inputs (adapted from Neural Networks Tutorial, 2017)**

[17] Andy Thomas, "Neural Networks Tutorial - A Pathway to Deep Learning," *Adventures in Machine Learning*, March 18, 2017, http://Adventuresinmachinelearning.com/neural-networks-tutorial/. (accessed July 29, 2018).

The node receives the inputs (X1, X2, X3…Xn). Each one is multiplied by its corresponding weight represented by (w1, w2, w3….) and then pulled into the activation function inside the node. These weights are numbers that change during the learning process. The output of the activation functions is called *F* in figure 4.

- Bias. An additional element is called *bias,* denoted by *b,* included in the below equation, and it has the job of rising (positive) or reducing (negative) the input of the activation function.

$$v = \sum_{j=1}^{n} wj.Xj + b \qquad (4)$$

Finally a version of a fully interconnected neural network with multiple layers is exhibit in the following diagram:

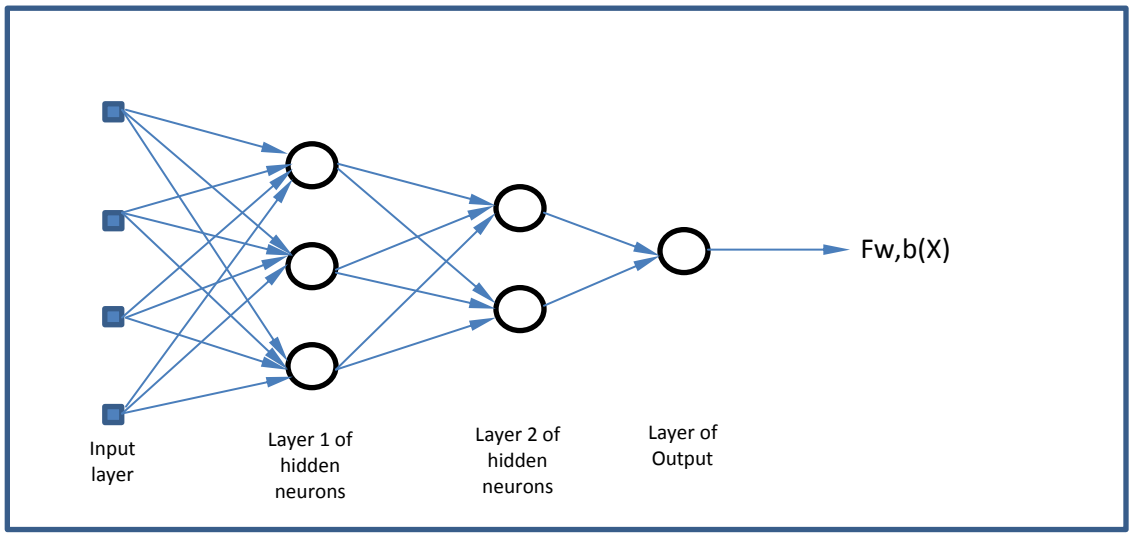



**Figure 5. Fully connected neural network**

In Figure 5, there are only two hidden layers and five neurons between the input and output but this setting is adjustable according to specific requirements.

Neural networks will be utilized in this dissertation as a tool in an attempt for identifying an algorithm capable of predicting hotel room prices, based on

acknowledged numerical attributes. These attributes will be divided in two groups. One, equivalent to the 80% of the data available, is called *data training* and will be uploaded into the input layer thus the network can learn from this data eventually finding patterns and generating a model. The reliability of this model will be put under test by predicting the remaining 20% of the data set. The model will then be assessed in terms of accuracy.

## 3.4 Decision tree

It is a graphical representation of a set of instances usually presented as a numerical or categorical attributes. A decision tree scans the whole dataset to find the most critical attribute to start with the split of data. This mapping resembles a tree where the root is the most important feature that splits its value through multiple nodes which in turn are conducting tests whose outcome is applied to build sub trees as figure 6 shows.[18] This process continues until a specific condition is reached usually associated with a leaf or target variable.

Decision trees are considered classifiers as the algorithm creates a model to predict the label for unseen instances. To achieve this, the classifier requires training data that is clearly separated from the prediction stage.[19]

The training data is used as an example by the algorithm which makes a prediction based on the current model and its outcome is tested against the validating data to assess the accuracy of the model.

[18] Rebecca Njeri, "What is a Decision tree Algorithm?," *Medium*, September 3, 2017, https://medium.com/@SeattleDataGuy/what-is-a-decision-tree-algorithm-4531749d2a17 (accessed August 16, 2018).

[19] Albert Bifet et al., *Machine Learning for data Streams* (London: The MIT Press, 2017)

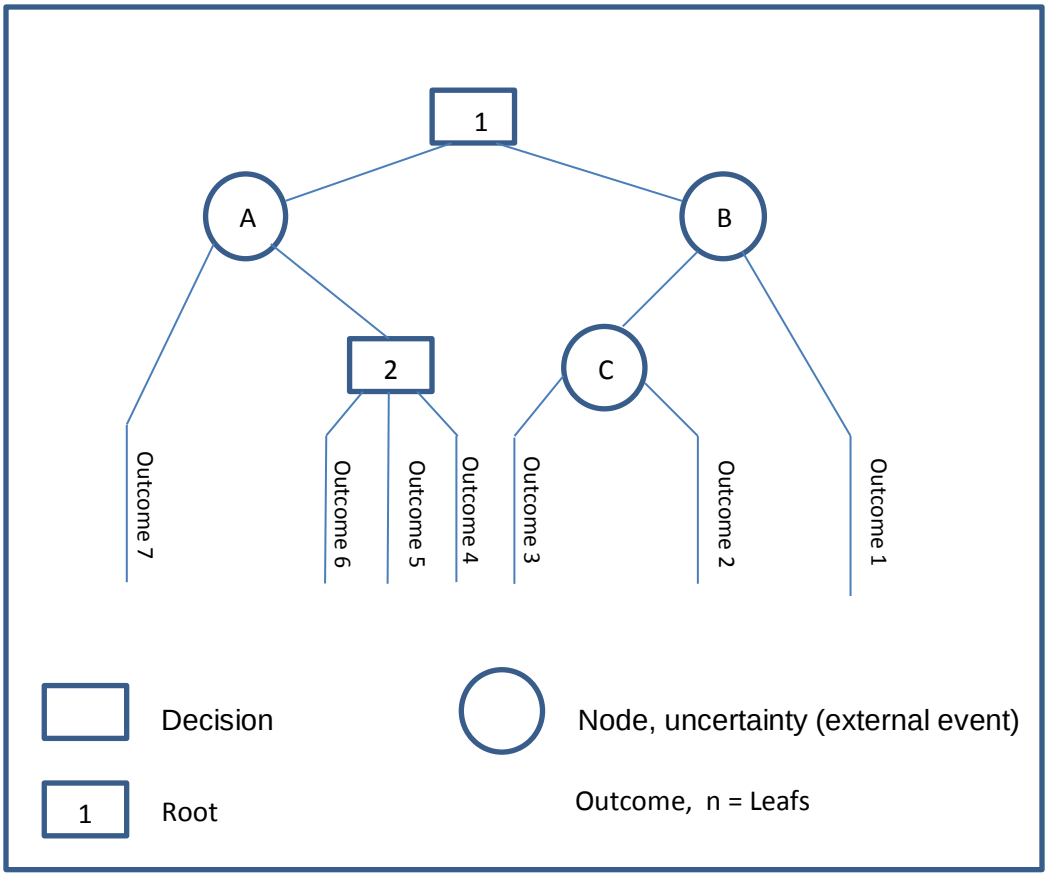



**Figure 6. Decision tree components**

Some advantages of decision trees include its easy interpretation and simplified data preparation process as not scaling is required. Its main downside is that it tends to get overfitted by too many parameters which might lead to an unreliable prediction.

# 4. DATA

## 4.1 Data Collection and sources

The data sets sourced contains information from five different hotel properties linked to the same hotelier chain located in the United States.[20] These properties represent a heterogeneous mixture in terms of size, price, location and customer's target as it can be observed below on the table 1:

| Hotel ID | Customer Segment (Income Level) | Location | Number of Rooms |
|---|---|---|---|
| 1 | High | Urban/Downtown | 670 |
| 2 | Medium | Suburban/Roadside | 60 |
| 3 | Medium | Suburban/Airport | 160 |
| 4 | Medium | Highway/Roadside | 70 |
| 5 | Medium to High | Urban/Downtown | 260 |

**Table 1. Hotel features**

The hotel chain mainly serves transient travelers meaning those with short length of stay. By focusing on these types of properties, the amount of time required to collect transactional data is shortened given the reduced booking horizon (normally less than four weeks) prior to check in compared to leisure hotels where bookings can be made several months in advance.

Collection of data takes place in multiples instances along the whole sales process, hence the importance of explaining first how the latter works in a standard way.

Rooms and other products of the hotel are sold through three distribution channels; central reservation systems (CRS) run by the hotel, global distribution systems

[20] Mark Ferguson, Tudor Bodea, "Choice-Based Revenue Management: Data from a major Hotel Chain, *InformsPubsOnLine,* September 9, 2008, https://pubsonline.informs.org/doi/suppl/10.1287/msom.1080.0231 (accessed May 4, 2018).

(GDS) and the hotel's website (WEB). The CRS is accessed by the hotel's reservations personnel from different locations around the world thus it can also be known under the name; central reservation office (CRO). This process is exhibit in figure 7.

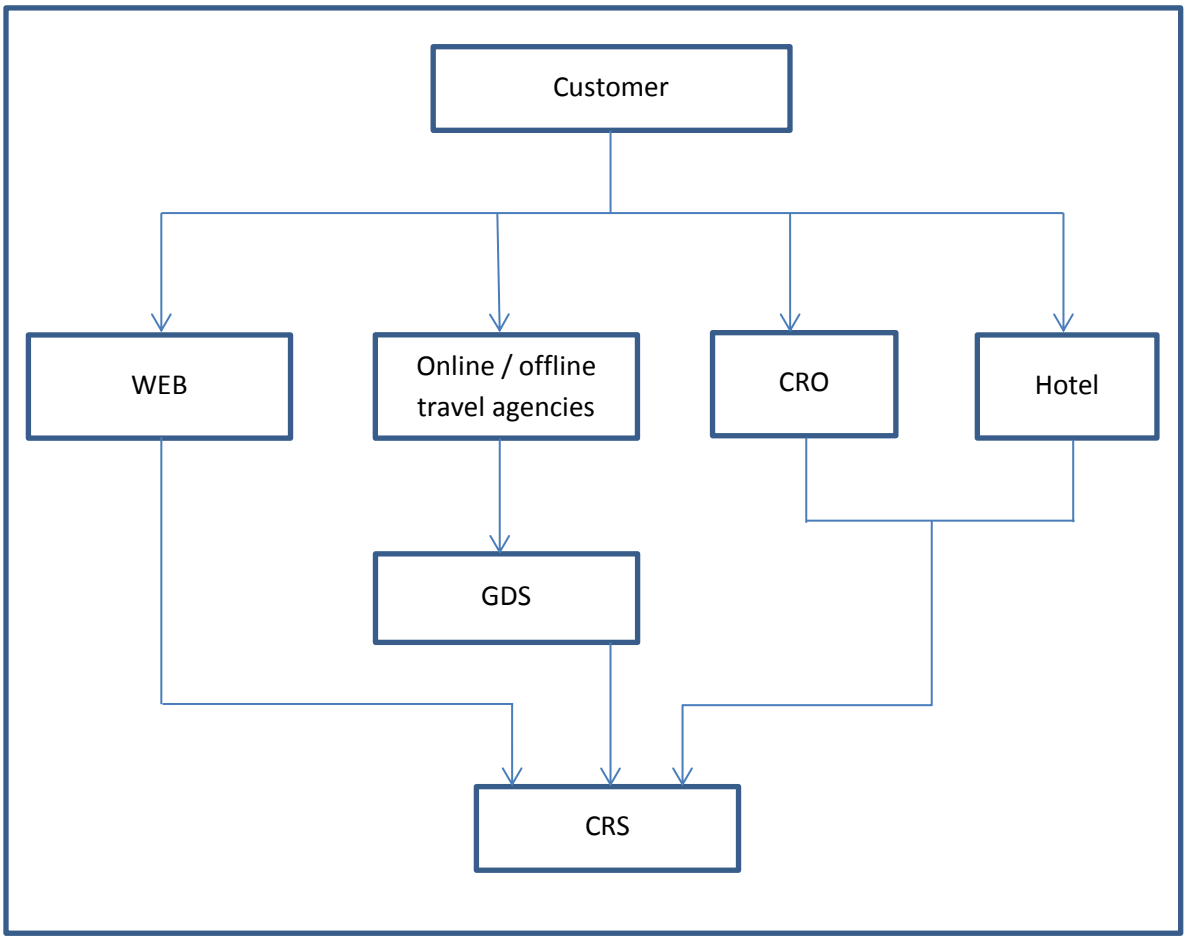



**Figure 7. Hotel selling process and distribution channels. (Adapted of Mark Ferguson, Tudor Bodea, 2008)**

CRO is a system that enables reservation agents to receive bookings requests over the phone and processing them immediately given the real time information sourced by CRS of the hotel. Unlike hotel's front desk or reservation departments, CRO works similar to a 'call center' and provide information concerning any of the hotel's properties.

Additionally, travel agencies require information in terms of room's availability, price, terms and conditions etc. This type of information is retrieved directly from the GDS servers which are updated by CRS.

All the information regarding bookings was collected between February 12th, 2007 and April 15th, 2007. As rates and room availability change continuously, a program based on Visual Basic script was employed to retrieve data in real time from the CRS. During the first two weeks, data was collected every six hours and the frequency increased to every 3 hours in the last two weeks.

Selling strategies are constantly changed and normally ruled by confidential polices within the organizations. That hinders the collection of information regarding availability of products and the way they are allocated amongst distribution channels. Hotels can restrict the type/quantity of rooms, rates or packages distributed by other sales parties in an attempt to increase profitability[21].

In order to collect data from travel agents, researches used a software tool called (ActivateState Software Inc. 2007) to verify the product availability against the reservation system run by the hotel. By conducting this query, was possible to determine which rooms and packages were not sold directly by the hotel, but by third parties. This procedure was supported with a manual validation process which took place on weekly basis.

On the other hand, information that is not covered in the data set object of this study includes: cancellations or no-shows, bookings made through on line travel agencies and special reservations derived of agreements with companies or government groups.

[21] Ibid.

## 4.2 Description of data set

For this study, a total of five datasets (one for each hotel) in comma separate values CSV format were downloaded[22]. Each data set contains transactional data and all five establishments belong to the same hotel chain.

The original data set with 25 variables had to be adjusted for the purpose of this exploratory analysis. The final data file called "*hotel*" exported into RStudio contains a total of 29 numerical and non-numerical attributes including 4346 observations. Variables or attributes are interchangeable terms in this study.

A description regarding the type of data utilized in this analysis is compressed in four measurement scales: nominal, ordinal, interval and ratio[23].

- Nominal. Its values represent categories without quantitative value or numerical significance. i.e Gender M = Male, F=Female
- Ordinal. The order of the values is significant and important for ranking. i.e Levels of satisfaction on a survey
- Interval. Also known as scale, their values depict categories in which the order is known as well as the distance between those values. i.e Differences in time and ranges of prices or income.
- Ratio. It is a scale that uses the value zero (0) meaning a total nonexistence of what is being measure. i.e The number of rooms sold in a specific hour.

Metrical data spans over numbers such as decimals, integers or cardinal. The latter are positive integers used to count things.

[22] Tudor Bodea, Mark Ferguson, Laurie Garrow, "Data Set – Choice Base Revenue Management: Data from a Major Hotel Chain," *informsPubsOnline*. Columbia Business School, September 9, 2008, https://pubsonline.informs.org/doi/suppl/10.1287/msom.1080.0231 (accessed April 4, 2018).

[23] Priya Chetti, Sudeshna, "Nominal, Ordinal and Scale in SPSS," *Project Guru,* January 16, 2015, https://www.projectguru.in/publications/nominal-ordinal-scale-spss/ (accessed August 11, 2018).

Rows from the data set were filtered according to the following criteria:
Purchased_Product labeled with one (1) represent the room and rate selected by the customer. Consequently, all the other products listed in the same column under the same booking number are disregarded. As an example see table 2.

| Booking_ID | Product_ID | Purchased_Product |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 0 |
| 2 | 2 | 0 |
| 2 | 3 | 0 |
| 2 | 4 | 0 |
| 2 | 5 | 0 |
| 2 | 6 | 0 |
| 2 | 7 | 0 |
| 2 | 8 | 1 |

**Table 2. Example of filtered list**

In this example, the first row at the top shows a list of one product (Product_ID) that was presented to the Booking ID number 1. This customer selected the only alternative available thus the column Purchased_Product was filled in with number 1.

For the Booking_ID number 2, a list containing eight different products was presented. They are vertically displayed in Product_ID column. The customer selected the number 8, so that, the number 1 was placed in the Purchase_Product column in the same row than the selected product. All rows marked with zero (0) in the Purchased_Product column are deleted.

The same procedure takes place for all five data sets and the reason for applying this methodology is that the analysis focuses only on those observations where a selection of a product was finally made avoiding unnecessary redundancies. The second adjustment takes place in columns containing dates. As the original

format was not suitable for additional calculations with the software, it was required to convert it to a standard dd/mm/yyyy format.

A third change occurs by re-calculating Advance_Purchase and Length_of_Stay columns. As a general rule applied; Advance_Purchase is the difference between Check_In_date and Booking_Date. Similarly, Length_of_Stay is the difference between Check_Out_Date and Check_In_Date.

The standard procedure followed to make sure that data will be suitable for a further examination exhibits this sequence:

1. Unification. All assigned variables are standard and valid for each hotel
2. Coding. Allocation of categories or values representing attributes to obtain valid measurements
3. Scaling. implementing a mechanism of normalization to enable each attribute contribute with similar proportion to the measurement
4. Merging. All five datasets are merged into one single table

After filtering the list, eight categories were identified for Purchased_Rate_Code. Consequently, two columns were derived from the original. One is containing the category or factor and the other a description. See table 3.

| Rate_Number | Rate_Number_Description |
|---|---|
| 1 | Advance Purchase. Requires at least 7 day Advance Purchase; Fully Restricted Rate |
| 2 | Rack Rate - Unrestricted Rate |
| 3 | Rack Rate Combined with Additional Hotel Services |
| 4 | Discount Rate Less Restricted Than Advance Purchase |
| 5 | Accommodation Combined with Frequent Traveler Rewards |
| 6 | Accommodation Combined with In Room Services |
| 7 | Accommodation Combined with Airport/Airline Services |
| 8 | Accommodation Combined with City / Weekend Activities |

**Table 3. Rate type category**

Purchase_Room_Type had originally 29 varieties which were impractical for the analysis whereby the necessity for clustering according to its essential use. The resulting categories count six different types of rooms as illustrated in table 4.

| Room_Type_Number | Room_Type_Number_Description |
|---|---|
| 1 | Double Beds Room |
| 2 | King Room |
| 3 | Suits |
| 4 | Special Type Room 1 |
| 5 | Queen Rooms |
| 6 | Standard Room |

**Table 4. Room type category**

Distribution_Channel column in table 5, has three different types so a nominal value has been allocated to each one of the categories.

| Distribution_Channel_Number | Distribution_Channel |
|---|---|
| 1 | CRO/Hotel |
| 2 | GDS |
| 3 | WEB |

**Table 5. Distribution channel category**

An interval was applied For Nightly_Rate attribute as shown in table 6 below:

| Nightly_Rate_Category | Nightly_Rate |
|---|---|
| 1 | More than 0 <= 80 |
| 2 | More than 80 <=150 |
| 3 | More than 150 <= 300 |
| 4 | More than 300 |

**Table 6. Nightly rate category**

Party_Size, exhibit in table 7, was categorized in the following way:

| Party_Size_Category | Party_Size | Party_Size_Description |
|---|---|---|
| 1 | 1 | Single |
| 2 | 2 | Double |
| 3 | 3 | Triple |
| 4 | 4,5,6,7,8,9,10,11,12 | Quad |

**Table 7. Party size category**

Based on the information provided in table 1, additional columns were incorporated to the dataset together with their own categories.

| Hotel_Location_Category | Hotel_Location |
|---|---|
| 1 | Urban / Downtown |
| 2 | Suburban / Roadside |
| 3 | Suburban / Airport |
| 4 | Highway / Roadside |

**Table 8. Hotel location category**

| Hotel_Size_Category | Number_Of_Rooms | Hotel_Size |
|---|---|---|
| 1 | 0 - 70 | Small |
| 2 | 71 - 259 | Medium |
| 3 | 260 o more | Big |

**Table 9. Rate Number**

| Income_Level_Category | Customer_Income_Level |
|---|---|
| 1 | Medium |
| 2 | Medium to High |
| 3 | High |

**Table 10. Rate Number**

The following columns were omitted from the original dataset as their significance to the current technical approach is not relevant or it can be explained by another existing attribute:

- Enrollment Date
- VIP Enrollment Date
- Number of Rooms
- Room Type
- Rate Code
- Arrival Rate
- Booking ID
- Purchased Product

The resulting list is made up of 29 variables as is shown in the table 11.

7 variables are denoted as integer, 18 categorical, 1 interval and 3 are dates.

| Variable / Column | Description | Type of variable |
|---|---|---|
| Hotel_ID | It indicates the number of the property labeled 1 to 5 | Categorical |
| Hotel_Location | Area where the property is located | Categorical |
| Hotel_Location_Category | Value indicating Hotel_Location | Categorical |
| Number_Of_Rooms | Amount of rooms per type of property | Integer |
| Hotel_Size | Description of the hotel size | Categorical |
| Hotel_Size_Category | Value indicating hotel size | Categorical |
| Product_ID | Represent a product available at the time of booking. The product is a combination of room type and rate | Integer |
| Customer_Income_Level | Description of customer according to their income level | Categorical |
| Income_Level_Category | Value indicating customer income level | Categorical |
| Booking_Date | It refers to the date when the booking was created | Date |
| Check_In_date | Date when the guest arrives at the hotel and is registered | Date |
| Check_Out_date | Date when the guest leaves the hotel and clears the bill | Date |
| Distribution_Channel_Number | Value indicating distribution channel | Categorical |
| Distribution_Channel | GDS, CRO/Hotel, WEB | Categorical |
| Advance_Purchase | Number of days prior to the check in date | Integer |
| Party_Size | Number of guests associated with the reservation | Integer |
| Party_Size_Description | Description of group of guests | Categorical |

| | | |
|---|---|---|
| Party_Size_Category | Values associated to the group of guests | Categorical |
| Length_of_Stay | Number of nights booked by the guest | Integer |
| Nightly_Rate_Category | Value indicating the nightly rate | Interval |
| Nightly_Rate | Rate per night of stay | Integer |
| Total_Revenue | Product of nightly rate times number of nights booked | Integer |
| Rate_Number | Code that identifies the purchased rate or rate number description | Categorical |
| Rate_Number_Description | Description of the rate in terms of type of room, price and additional services | Categorical |
| Room_Type_Number | Code that explains the room type description | Categorical |
| Room_Type_Number_Des | Type of room booked by the guest | Categorical |
| Merge_Indicator | When purchase rate code cannot be matched at the time of the check in, it is denoted by zero | Categorical |
| Membership_Status | It equals 1 if guest has a membership. Otherwise 0 | Categorical |
| VIP Membership Status | It equals 1 if the VIP guest with Rewards Program Member has a membership status | Categorical |

**Table 11. Description of variables**

# 5. ANALYSIS OF DATA

## 5.1 Introduction

This chapter contains a description of the procedure followed to analyze the data set for each of the models stated on chapter 3. All three models are run on a software called RStudio which is a powerful open source programing language capable of performing complex calculations and statistical tasks. It is highly customizable and its versatility allows users upload data sets in multiple formats, also featuring an interface that displays graphs and windows for coding. This software is widely used across several fields including science, education and industry.

The first model explained in the Section 5.2, looks for correlations amongst 6 numerical and 13 categorical variables. Then, the numerical data will be prepared to fit a neural network, described in Section 5.3. And finally, a decision tree model, handling both numerical and text attributes, is reviewed in Section 5.4.

## 5.2 Correlations

The *hotel* data set is uploaded into RStudio by using the following command:

```
hotel<- read.csv(file.choose(), header = T, sep=";")
head(hotel)
```

It comprises the name of the file *hotel,* a designated function called *read.csv ()* which specifies the type of format that the software should look for, *file.choose()* to identify the file location and lastly, *header* and *sep* to indicate the presence of a row containing a header and semicolon separating the columns respectively.

In order to obtain the coefficient correlation, the imported file requires to be filtered excluding variables denoted by text. To do so, the following script was executed:

```
n_variable <- hotel_filtered[, c(2,9,10,12,14)]
```

The new data set named *n_variable* contains 5 numerical variables. They are listed below and have been described in the preceding chapter:

- Number_of_Rooms
- Advance_Purchase_Adjusted
- Party_Size
- Length_of_Stay_Adjusted
- Nigtlhy_Rate

The function to calculate the correlation Pearson is denoted by *cor ( )* as shown below:

```
hotel_cor <- cor(n_variable, method = "pearson")
round(hotel_cor,2)
```

To facilitate the lecture and tabulating results, two decimals were set per attribute.

When non-numerical variables or factors are being subject of analysis for correlations, then the Pearson coefficient method is not applicable. In this case, factors are paired and displayed in a chart called contingency table that shows its distribution and weighing.

Bar plots, as a visual tool, are also utilized to facilitate the interpretation of the contingency table outcome. Finally, a Chi-Squared test with significance level set to 0.05 is executed to determine whether the two assessed variables are independent or non-independent.

## 5.3 Neural Networks

The goal when using this method of machine learning is to find a valid model through which the nightly price per room is explained by the other numerical variables. By following an identical sequence of steps observed in the previous Section, the dataset is uploaded, filtered and ready to use in RStudio.

Consequently with the objective, price per room has been denoted as dependent and all the other numerical attributers constitute independent variables. An additional software package known as a "neuralnet" especially designed for processing neural networks is downloaded and installed into RStudio. The numerical variables used with this model are:

- Number_of_Rooms
- Product_ID
- Advance_Purchase_Adjusted
- Party_Size
- Length_of_Stay_Adjusted
- Nightly_Rate (dependent variable)

Random sampling and partition becomes the first step for processing data. In this stage *hotel_filtered* data set is randomly divided into two groups. 80% is a sample assigned to training and the remaining 20% becomes the testing part of the data set. These two groups are equivalent to 3476 and 870 observations respectively. The necessary code is presented as follows:

```
hotel_int <- hotel[, c(4,7,15,16,19,21)]
head(hotel_int)

## Random Sampling ##

samplesize = 0.80 * nrow(hotel_int)
set.seed(80)
index<-sample(seq_len(nrow(hotel_int)), size = samplesize)
```

Sampling in RStudio is performed using the *sample ( )* function. *set.seed ( )* is used to generate random sample and maintain consistency. It is also required, while fitting the neural network, to have indexed the two samples of data as illustrated in the script with the name *index*. For quality verification of the prediction, training and testing sets are created using the script:

```
datatrain = hotel_set[ index, ]
datatest = hotel_set[ -index, ]
```

It should be noted that the data set containing the training sample is now called "datatrain" and similarly the sample data for testing purposes is labeled "datatest".

Scaling the dataset is the next task. By doing so, a negative impact on the prediction variable derived of having attributes with large values is prevented. The difference between a scaled and non-scaled datasets is illustrated with the help of histograms in figure 8.

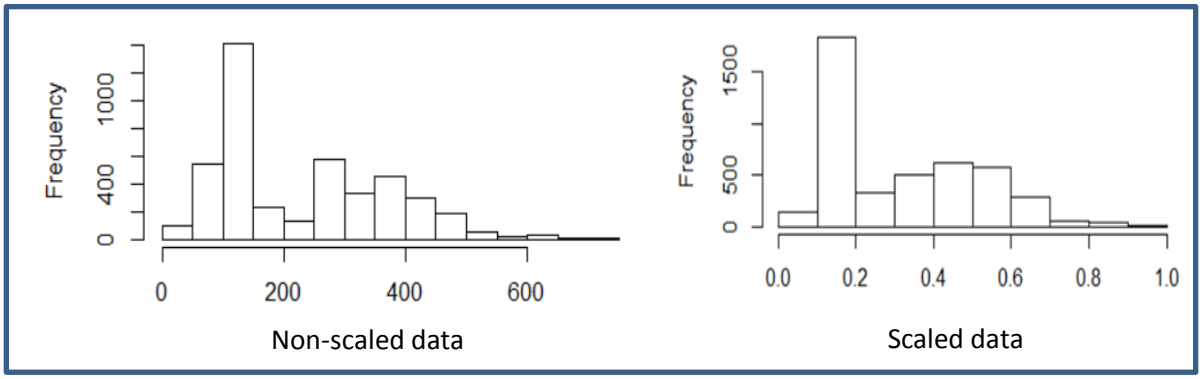



**Figure 8. Histograms with scaled an non-scaled data**

The non-scaled histogram shows how huge disparity in value among independent variables (bars) is perceived. This wide gaps bias the way an algorithm calculates the distance between two points. Therefore, it is necessary to implement a mechanism of normalization to enable each attribute contribute with similar proportion to the final distance. The histogram on the right, exhibits a normalized

scaled dataset. Distances have been kept proportionally and the difference among values turns into a more uniform shape in a range from 0 to 1.

For this dissertation a method known as *min-max* has been selected for data normalization as it preserves the initial distribution of the attributes.[24] The corresponded formula is stated below:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (5)$$

In *min-max* formula, $x'$ is the scaled value ranging between 0 and 1. The variable $x$ stands for the original value of the attribute and min and max are the lowest or maximum values respectively found in the dataset.

The script performed in R for scaling the data is:

```
max = apply(hotel_set , 2 , max)
min = apply(hotel_set , 2 , min)
scaled = as.data.frame(scale(hotel_set, center = min, scale = max - min))
```

Now, one dataset for training and one for testing are created by using the scaled data frame obtained in the previous step.

```
trainNN = scaled[index , ]
testNN = scaled[-index , ]
```

The neural network is fitted with the *trainNN* dataset. It should be observed that Nightly Rate as dependent variable is explained by the other 5 numerical variables and the notations in R is as follows:

[24] Chaitanya Sagar, "Creating and Visualizing Neural Network in R," *Analytics Vidhya*, September 7, 2017, https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/. (accessed July 29, 2018).

```
set.seed(2)

NN <- neuralnet(trainNN$Nightly_Rate~ trainNN$Number_Of_Rooms +
                                      trainNN$Product_ID +
                                      trainNN$Advance_Purchase_Adjusted +
                                      trainNN$Party_Size +
                                      trainNN$Length_of_Stay_Adjusted,
                                      data = trainNN,
                                      hidden = c(2,1),
                                      act.fct = "logistic",
                                      linear.output = TRUE)
```

The hidden argument indicates the number of neurons behind each hidden layer, while the argument *linear.output*, means that linear regression is required by the user.[25]

As a next step, the prediction addressing the scaled testing data is performed employing the function *compute ( )*. The number 6 corresponds to the column where the Nighty Rate is located in the dataset as illustrated in the coding below:

```
predict_testNN = compute(NN, testNN[,c(-6)])
```

All variables in the prediction set remains normalized hence the need to scaling them back to a standard values for a meaningful comparison. The coding is as follows:

```
predict_testNN = (predict_testNN$net.result * (max(hotel_int$Nightly_Rate)-
                                               min(hotel_int$Nightly_Rate)))+
                                               min(hotel_int$Nightly_Rate)
```

The comparison between the prediction using unscaled data and the validation data is an important step to visualize the accuracy of the model. Both sets are placed on a same plot using this script:

[25] Michy Alice, "Fitting a Neural Network in R; neuralnet package," *datascience+*, September 23, 2015, https://datascienceplus.com/fitting-neural-network-in-r/. (accessed July 28, 2018).

```
plot(datatest$Nightly_Rate~ predict_testNN, col='blue',
                                            pch=16,
                                            cex.axis = 1.0,
                                            main = "Prediction Quality vizualizat
                                            ylab = "predicted Nightly Rate",
                                            xlab = "Real Nightly Rate")
```

From a statistical point of view, the error between the two variables (observed and predicted nightly rates), is calculated by using a function that measures their dispersion. This method known as a mean absolute error (MAE) is depicted with the formula:

$$\text{MAE} = \frac{\sum_{i=1}^{n} | y_i - x_i |}{n} \qquad (6)$$

It takes the absolute difference between the variables above commented and determines the average.[26] The way RStudio performs this function is expressed in the following script:

```
err_predict_NN = mean(abs(predict_testNN - datatest$Nightly_Rate))
```

The terms *mean* or *averages*, are equivalent and *abs ( )* is a function that computes the absolute value of the numerical variable resulting from the subtraction amid the prediction and the observed testing data. Once the error has been found, it requires being contrasted with its similar from an alternative prediction method. For this analysis, and given the trend exhibited in the previous step, a linear regression is developed resulting in a similar path than the used with the neural network. The script for the linear model is the following:

[26] Ben Rogojan, "How to Measure the Accuracy of Predictive Models," *acheron analytics*, May 28, 2018, http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models#. (accessed August 01, 2018).

```
TRN_LM = scaled[index , ]

TST_LM = scaled[-index , ]


model = lm(Nightly_Rate ~ Number_Of_Rooms+
                          Product_ID +
                          Advance_Purchase_Adjusted +
                          Party_Size +
                          Length_of_Stay_Adjusted,
                          data = TRN_LM)


predict_testLM = predict(model, TST_LM[,c(-6)])
```

The output needs to be scaled back to its standard value as explained before and its error will be contrasted with the one obtained in the previous step to determine whether neural network approach presents more accuracy in the prediction that the alternative linear regression method or not.

## 5.4 Decision Tree

Unlike the preceding model, decision tree allows to use both numerical and categorical variables also known as factors. Thus, this model incorporates more attributes in an attempt to gain a greater inside on how they are influencing the setup of the dependent variable, nightly room price.

A software package builded in RStudio called "party" is used for this part of the analysis and the attributes are:

- Hotel_Location_Category Factor
- Number_of_Rooms Integer
- Hotel_Size Factor
- Product_ID Integer
- Customer_Income_Level Factor
- Distribution_Channel_ Factor

- Advance_Purchase_Adjusted — Integer
- Party_Size_Description — Factor
- Length_of_Stay_Adjusted — Integer
- Nightly_Rate_Category (dependent variable) — Factor
- Room_Type_number — Factor
- Rate_Number — Factor
- Membership_Status — Factor

After data has been imported, is necessary let RStudio know the type of variable present in the dataset. To do this the following script has been written:

```
hotel_tree <- hotel[ , c(3,4,5,7,8,14,15,17,19,20,23,25,28)]
str(hotel_tree)
head(hotel_tree)

## Converting variables into the right classification

hotel_tree$Hotel_Location_Category<-as.factor(hotel_tree$Hotel_Location_Category)
hotel_tree$Nightly_Rate_Category<-as.factor(hotel_tree$Nightly_Rate_Category)
hotel_tree$Rate_number<-as.factor(hotel_tree$Rate_number)
hotel_tree$Room_Type_number<-as.factor(hotel_tree$Room_Type_number)
hotel_tree$Membership_Status<-as.factor(hotel_tree$Membership_Status)
```

Random sampling and partition are executed on *hotel_tree* data set and it is randomly divided into two groups. 80% assigned to training and the remaining 20% to validation or testing data. These two groups are equivalent to 3476 and 870 observations respectively. The necessary code is presented as follows:

```
# Random sampling

samplesize = 0.80 * nrow(hotel_tree)
set.seed(80)
index_tree <- sample(seq_len(nrow(hotel_tree)), size = samplesize)
```

In the next step, samples for training and other for validation are created by using the coding:

```
# Create training and test datasets

train_hotel = hotel_tree[index_tree, ]
test_hotel = hotel_tree[-index_tree, ]
```

The decision tree is fitted with the variables correctly classified and the train data as depicted in the script below:

```
# fit decision tree

h_tree <- ctree(Nightly_Rate_Category~ Hotel_Location_Category+
                  Number_Of_Rooms+
                  Hotel_Size+
                  Product_ID+
                  Customer_Income_Level+
                  Distribution_Channel+
                  Advance_Purchase_Adjusted+
                  Party_Size_Description+
                  Length_of_Stay_Adjusted+
                  Rate_number+
                  Room_Type_number+
                  Membership_Status,
                  data = train_hotel,
                  controls = ctree_control(mincriterion = 0.99, minsplit = 1000))
```

Notice that Nightly_Rate_Category is the dependent categorical variable defined by the others 12 attributes (numerical and factors) representing independent variables. In the last line of the script there is a function called, *controls,* utilized to trim the tree upon two parameters; *mincriterion* which sets the confidence level of significance for the variable and *minsplit* positioned to define when the algorithm should split into two a branch based on a certain sample size. The output shows a total of 15 nodes including all variables with its corresponding weighing calculated by the algorithm as presented below:

```
1) Customer_Income_Level == {Medium, Medium to High}; criterion = 1, statistic = 2878.255
  2) Hotel_Location_Category == {1, 2}; criterion = 1, statistic = 558.592
    3)*  weights = 497
  2) Hotel_Location_Category == {3, 4}
    4) Rate_number == {2, 5, 6, 7}; criterion = 1, statistic = 119.258
      5) Room_Type_number == {1, 2, 5, 6}; criterion = 1, statistic = 44.071
        6) Product_ID <= 7; criterion = 1, statistic = 25.559
          7)*  weights = 400
        6) Product_ID > 7
```

```
            8)*  weights = 606
        5) Room_Type_number == {3}
          9)*  weights = 30
      4) Rate_number == {1, 4, 8}
        10)*  weights = 370
P   1) Customer_Income_Level == {High}
e     11) Rate_number == {2, 3, 5}; criterion = 1, statistic = 136.513
        12) Advance_Purchase_Adjusted <= 43; criterion = 1, statistic = 66.779
          13)*  weights = 1036
        12) Advance_Purchase_Adjusted > 43
          14)*  weights = 210
      11) Rate_number == {1, 8}
        15)*  weights = 327
          predict_datatest_t
```

A visual way to represent both sets; predicted and testing, to verify the accuracy of the model is through tabulation. The following coding in R has been employed:

```
## Misclassification error for testing data ##

table_hotel2<- table(predict_datatest_t,test_hotel$Nightly_Rate_Category)

table_hotel2
```

The function *table ()* in R does the job by placing both sets in a single table with numerical data and a misclassification occurs when there is no coincidence between the predicted and validating data. This type of errors can be quantified by the following R script:

```
1-sum(diag(table_hotel2))/sum(table_hotel2)
```

The diagonal in the chart contains all coincidences, which are successes for the prediction and all the other numbers should be considered errors.

# 6. RESULTS

## 6.1 Introduction

This chapter will develop an overview of the main findings associated with the use of deep learning tools for the scientific approach to data analysis and the benefits derived of their utilization in revenue management within the hotel industry. Numerical and categorical variables which fundamentally are reflecting nuances of the business including seasonality factors, customer behavior, marketing, supply and demand, allow hotels to have variable room rates which in turn enable them to maximize revenue and occupancy rates[27]. In this context, the examination of data performed, aims to determine interdependencies among attributes highlighting reference values that will be analyzed in the following chapter. Section 6.3 looks into a neural network technique to trace an algorithm for price prediction and in section 6.4 the target of the study will be to implement a technic for classification and forecasting.

## 6.2 Correlations

Since attributes are mainly represented in this data set by categorical variables, which cannot be analyzed by using the Pearson coefficient method, they had been initially displayed in pairs within contingency tables to facilitate the interpretation

[27] altexsoft, **"**Machine Learning Redefines Revenue Management and Dynamic Pricing in Hotel Industry",altexsoft, 4 June 2018, https://www.altexsoft.com/blog/datascience/machine-learning-redefines-revenue-management-and-dynamic-pricing-in-hotel-industry/ (accessed 24 August, 2018)

of its association. A Chi-squared test is also part of the analysis to help to conclude whether the assessed attributes are independent or dependent of each other.

### 6.2.1 Nightly rate category and distribution channel

The first pair examined is nightly rate category and distribution channel. Nightly rates have been split in four categories and its distribution rates intervals are shown in figure 9.

Prices between 80 and 150 dollars, labeled with the number two were predominantly higher among the five hotels meaning that more than 40% of the sold rooms fell in this category followed by rooms with prices above 300 dollars.

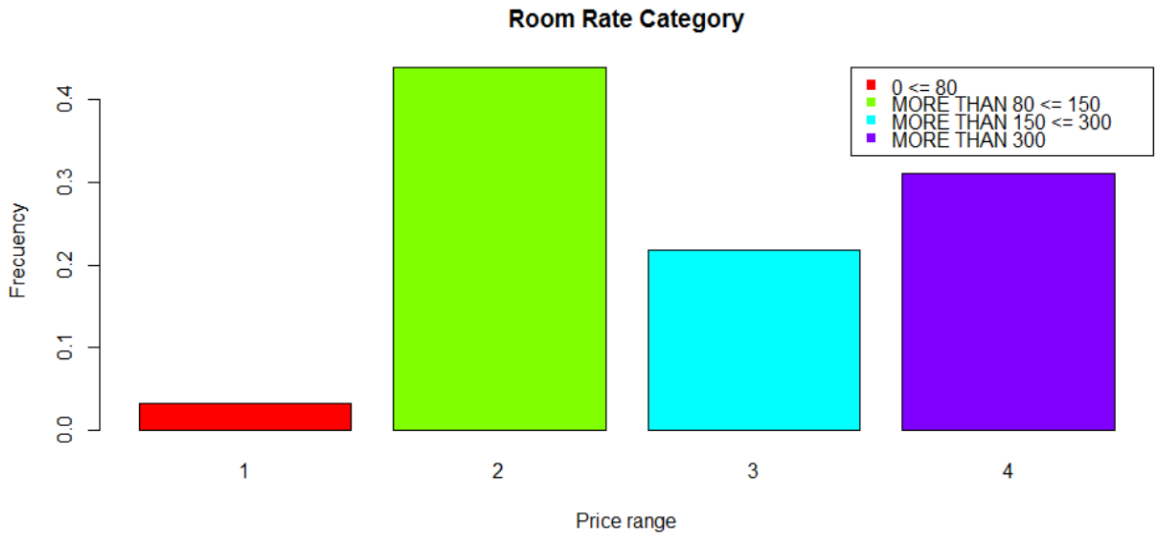



**Figure 9. Room price categories**

The other variable of this pair, distribution channel, represented in figure 10, shows that more than 60% of the bookings were made through CRO/Hotels. The contribution of Global Distribution Channels - GDS to the total sales did not even reached 10%.

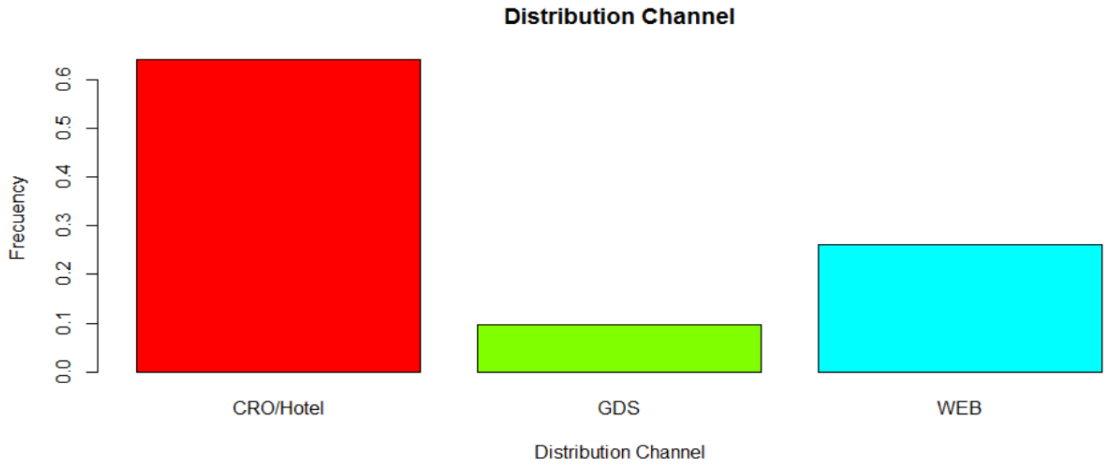



**Figure 10. Distribution channels**

After combining the results of the above two variables a contingency table 12, is generated. The table shows the distribution by price interval and distribution channel and it counts a total of 4346 bookings.

It can be observed that 31% of the reservations (largest contribution) equivalents to 1352 are located in the range price between 80 and 150 dollars, level 2, and they were made through the CRO/Hotel distribution Channel.

| | CRO/Hotel | | GDS | | WEB | | TOTAL | |
|---|---|---|---|---|---|---|---|---|
| 1 | 95 | 0.02 | 3 | 0.00 | 44 | 0.01 | 142 | 0.03 |
| 2 | 1352 | 0.31 | 91 | 0.02 | 464 | 0.11 | 1907 | 0.44 |
| 3 | 551 | 0.13 | 102 | 0.02 | 291 | 0.07 | 944 | 0.22 |
| 4 | 787 | 0.18 | 226 | 0.05 | 340 | 0.08 | 1353 | 0.31 |
| **TOTAL** | 2785 | 0.64 | 422 | 0.09 | 1139 | 0.27 | 4346 | 1 |

**Table 12. Contingency table. Distribution Channel and Nightly rates**

Continuing with the analysis, this time visualizing figure 11 below, the lowest contribution to the revenue was derived from those rooms sold under 80 dollars. Particularly in GDS, the majority of sales corresponded to category four with rooms rates soaring over 300 dollars per night.

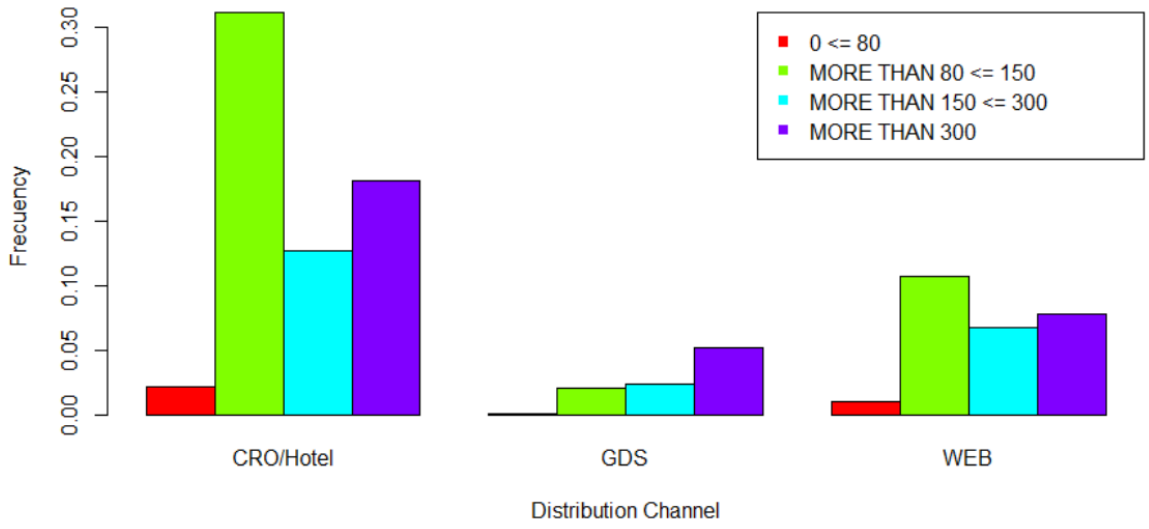



**Figure 11. Distribution channel and nightly rates**

The analysis of this pair is completed by conducting a Chi-squared test with significant level set to 0.05, where;

Ho. Nightly rate and distribution channel are independent
Ha. Nightly rate and distribution channel are dependent

```
        Pearson's Chi-squared test

data:  test_table
X-squared = 163.92, df = 6, p-value < 2.2e-16
```

As the significant level is less than 0.05 according to the p-value, then Ho is rejected and is concluded with 95% of confidence that nightly rate and distribution channel are dependent.

### 6.2.2 Hotel size and distribution channel

Hotel size has been divided in three categories small, medium and big and according to figure 12, more than 50% of the bookings equivalent to 2270 of this sample, were covered by big hotels with more than 260 rooms. A reduced contribution of 15% approximately of bookings or 630 came from small size properties.


**Figure 12. Distribution of bookings by hotel size**

When both sets of variables are merged, the result looks like the shown in table 13. CRO/Hotel is the main distribution channel among all the hotel sizes with 64% of the total bookings followed by WEB and GDS with 26% and 10% respectively.

| | BIG | | MEDIUM | | SMALL | | TOTAL | |
|---|---|---|---|---|---|---|---|---|
| **CRO/HOTEL** | 1301 | 0.30 | 1063 | 0.24 | 421 | 0.10 | 2785 | 0.64 |
| **GDS** | 329 | 0.08 | 62 | 0.01 | 31 | 0.01 | 422 | 0.10 |
| **WEB** | 640 | 0.15 | 321 | 0.07 | 178 | 0.04 | 1139 | 0.26 |
| **TOTAL** | 2270 | 0.53 | 1446 | 0.32 | 630 | 0.15 | 4346 | 1 |

**Table 13. Contingency table. Hotel size and distribution channels**

It can also be observed in table 13 that big and medium size hotel properties concentrated 85% of the total reservations.

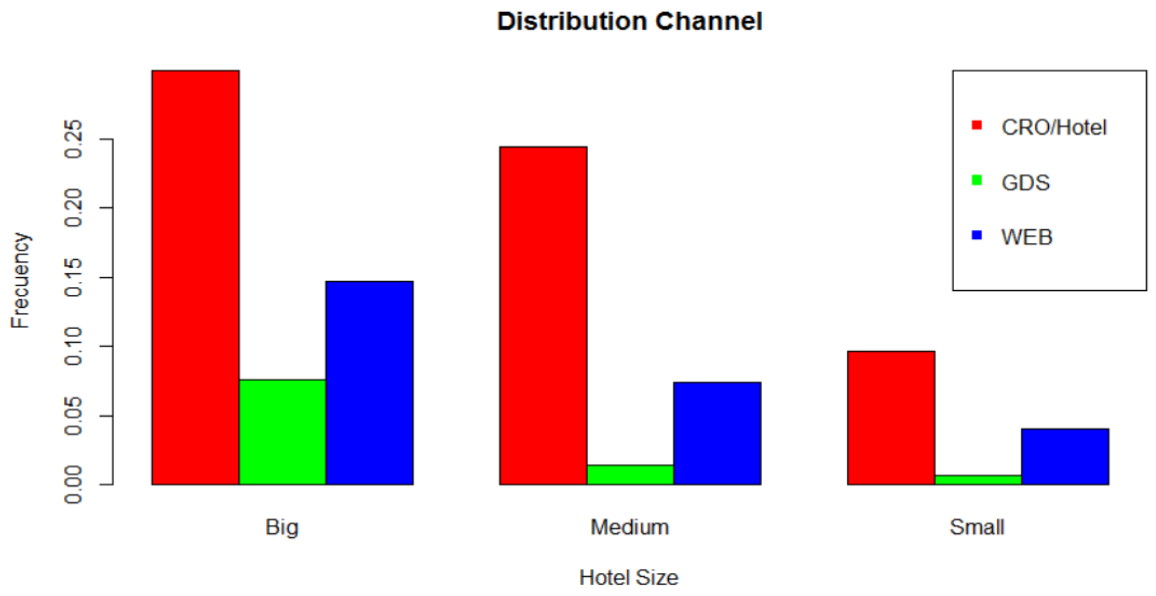



**Figure 13. Distribution of bookings by hotel size and distribution channel**

To determine dependency a Chi-squared test with significant level set to 0.05 is executed, where;

Ho. Hotel size and distribution channel are independent

Ha. Hotel size and distribution channel are dependent

```
        Pearson's Chi-squared test

data:  DC_HS
X-squared = 162.49, df = 4, p-value < 2.2e-16
```

As the significant level is less than 0.05 according to the p-value, then Ho is rejected and is concluded with 95% of confidence that hotel size and distribution channel are dependent.

### 6.2.3 Nightly rate category and rate number

Nightly rate category has been analyzed previously in figure 9, So that, the remaining variable for this pair, rate number, can be associated with the type of booking purchased and is represented in figure 14.


**Figure 14. Distribution of bookings by rate number**

The most sold type of room with 57% of the total bookings, belongs to the category 2, rack rate unrestricted - rate followed by discount rate with 15%. As for the type of booking with less preference among customers, accommodation with in room services is found with only 1% of contribution.

The contingency table 14 is displayed below with data summarizing distribution of both examined variables. In the first column on the left, all eight rate categories are vertically listed as shown in figure 14, and in the first row on the top, the numbers one to four stands for the price category denoted in figure 9. Regarding rate number 2 (rack rate – unrestricted rate), it had significant contribution especially when room prices were situated in level 2, that is, between 150 and 300

dollars. It should also be highlighted that under the same rate level 2 above mentioned it was sold the highest amount of accommodation combined with airport/airline services, rate number 7, equivalent to (6%) of the total room sales.

| | 1 | | 2 | | 3 | | 4 | | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 23 | 0.01 | 137 | 0.03 | 176 | 0.04 | 131 | 0.03 | 467 | 0.11 |
| **2** | 49 | 0.01 | 1000 | 0.23 | 491 | 0.11 | 949 | 0.22 | 2489 | 0.57 |
| **3** | 0 | 0.00 | 6 | 0.00 | 40 | 0.01 | 209 | 0.05 | 255 | 0.06 |
| **4** | 38 | 0.01 | 464 | 0.11 | 115 | 0.03 | 0 | 0.00 | 617 | 0.15 |
| **5** | 0 | 0.00 | 32 | 0.01 | 20 | 0.00 | 6 | 0.00 | 58 | 0.01 |
| **6** | 0 | 0.00 | 5 | 0.00 | 1 | 0.00 | 0 | 0.00 | 6 | 0.00 |
| **7** | 24 | 0.01 | 263 | 0.06 | 19 | 0.00 | 0 | 0.00 | 306 | 0.07 |
| **8** | 8 | 0.00 | 0 | 0.00 | 82 | 0.02 | 58 | 0.01 | 148 | 0.03 |
| **TOTAL** | 142 | 0.04 | 1907 | 0.44 | 944 | 0.21 | 1353 | 0.31 | 4346 | 1 |

**Table 14. Contingency table. Nightly rate category and rate number**


**Figure 15. Distribution of bookings by rate category and rate number**

At this point Chi-squared test with significant level of 0.05 is performed for these two variables as shown in the next code:

```
data:  DR_Number
X-squared = 1259.8, df = 21, p-value < 2.2e-16
```

Ho. Nightly rate category and type of rate are independent

Ha. Nightly rate category and type of rate are dependent

As the significant level is less than 0.05 according to the p-value, then Ho is rejected and is concluded with 95% of confidence that nightly rate category and type of rate are dependent.

### 6.2.4 Correlation of number of rooms, advance purchase, party size, length of stay, and nightly rate

This part of the analysis covers solely numerical variables thus correlation by using Pearson coefficient is applicable. The table 15, below comprises correlations of five numerical attributes assessed.

| | Number of rooms | Advance purchase | Party size | Length of stay | Nightly rate |
|---|---|---|---|---|---|
| **Number of rooms** | 1.00 | 0.25 | 0.07 | 0.14 | 0.87 |
| **Advance purchase** | 0.25 | 1.00 | 0.21 | 0.14 | 0.16 |
| **Party size** | 0.07 | 0.21 | 1.00 | 0.03 | 0.03 |
| **Length of stay** | 0.14 | 0.14 | 0.03 | 1.00 | 0.13 |
| **Nightly rate** | 0.87 | 0.16 | 0.03 | 0.13 | 1.00 |

**Table 15. Correlations applying Pearson coefficient method**

As show in table 15, a positive high correlation of 0.87 exists between nightly rate and number of rooms. It means that the larger the hotel measured by its number of rooms, the higher the rate charged to their guests.

The rest of the numerical attributes, subject to this type of testing, seem having neither strong association nor defined direction.

## 6.3 Neural network results

Numerical variables are also assessed by using neural network method. Figure 16 illustrates the outcome setup.

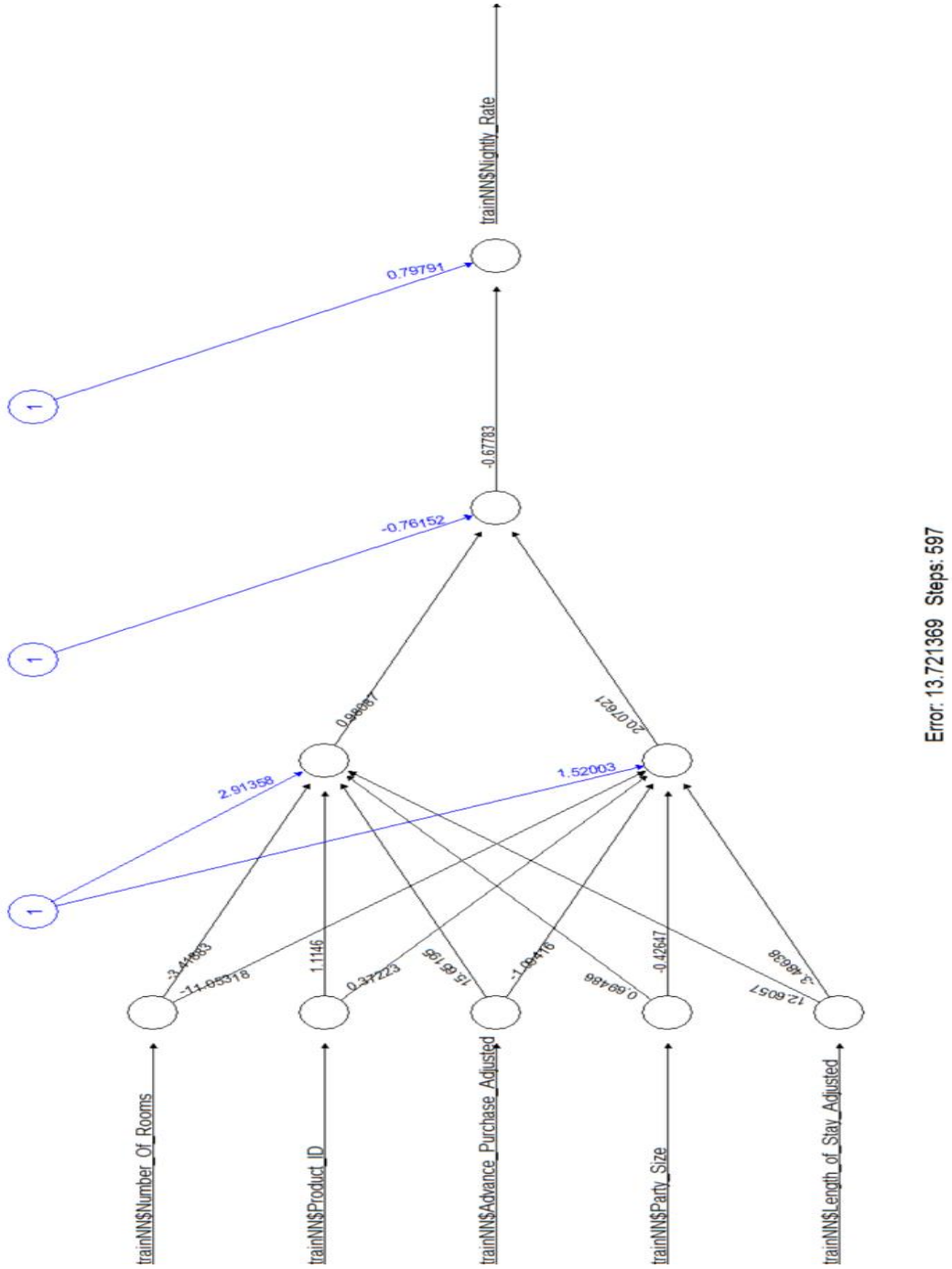



**Figure 16. Neural network with six numerical variables**

In the neural network in figure 16, the first hidden layer is made up of 2 neurons and the second hidden layer contains 1. It should also be observed, that every node or neuron is fully connected and fitted with five different types of variables coming from the training data set. The outcome is a unique node at the end with the predictive variable, nightly rate.

The resultant algorithm is fitted with the testing data set which excludes the variable, nightly rate, as it is supposed to be predicted. Two plots in figure 17 show the validated and predicted nightly rate data sets.

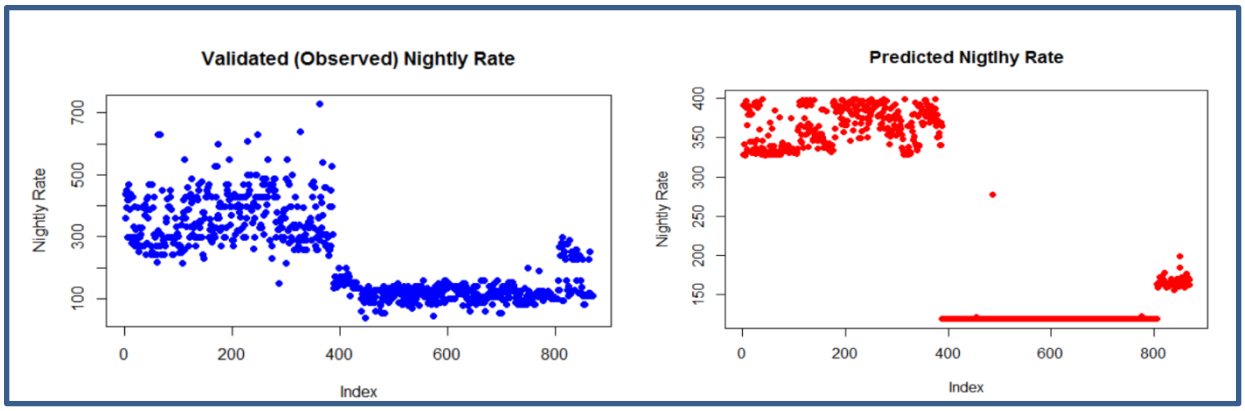



**Figure 17. Comparison. Validated and predicted data sets**

As illustrated in figure 17, the chart on the right containing predictive data, shows that the prediction works out fairly well especially for the first 400 observations indicated with the axis Index. Between the observations 400 to 800 the validation data set on the left, displays wider dispersion not captured by the algorithm in the prediction, which in turn, lessens its accuracy.

When both of the above data sets are placed in the same chart, the resulting plot is presented in figure 18. Ideally, all points should converge along the same trend line meaning a perfect prediction but this is not the case in most of the analysis. It is valid to point out that the algorithm has identified a clear pattern regarding nightly room prices though.

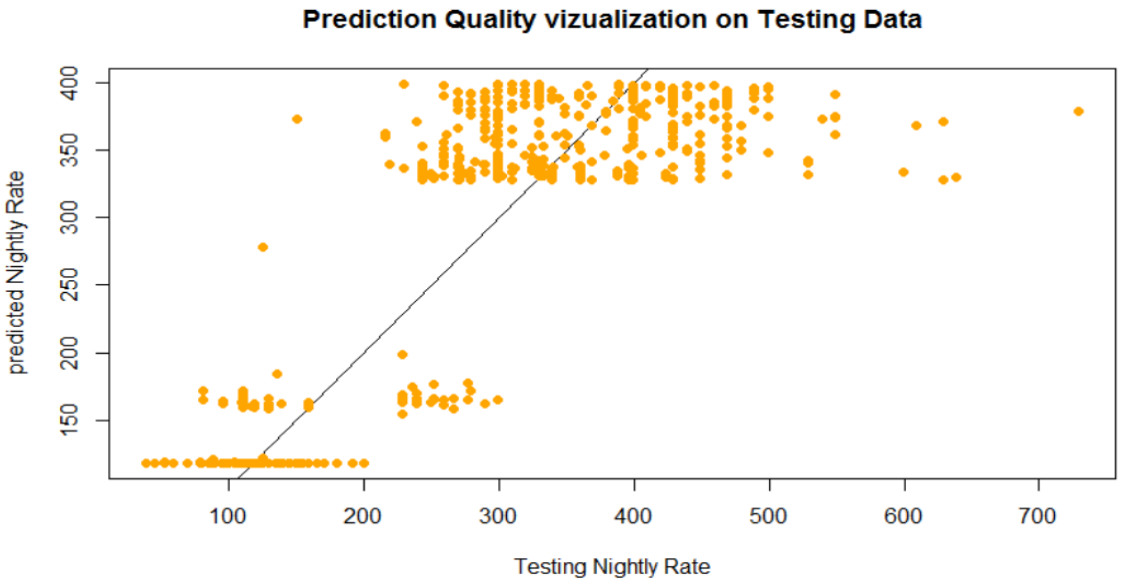



**Figure 18. Prediction quality on testing data**

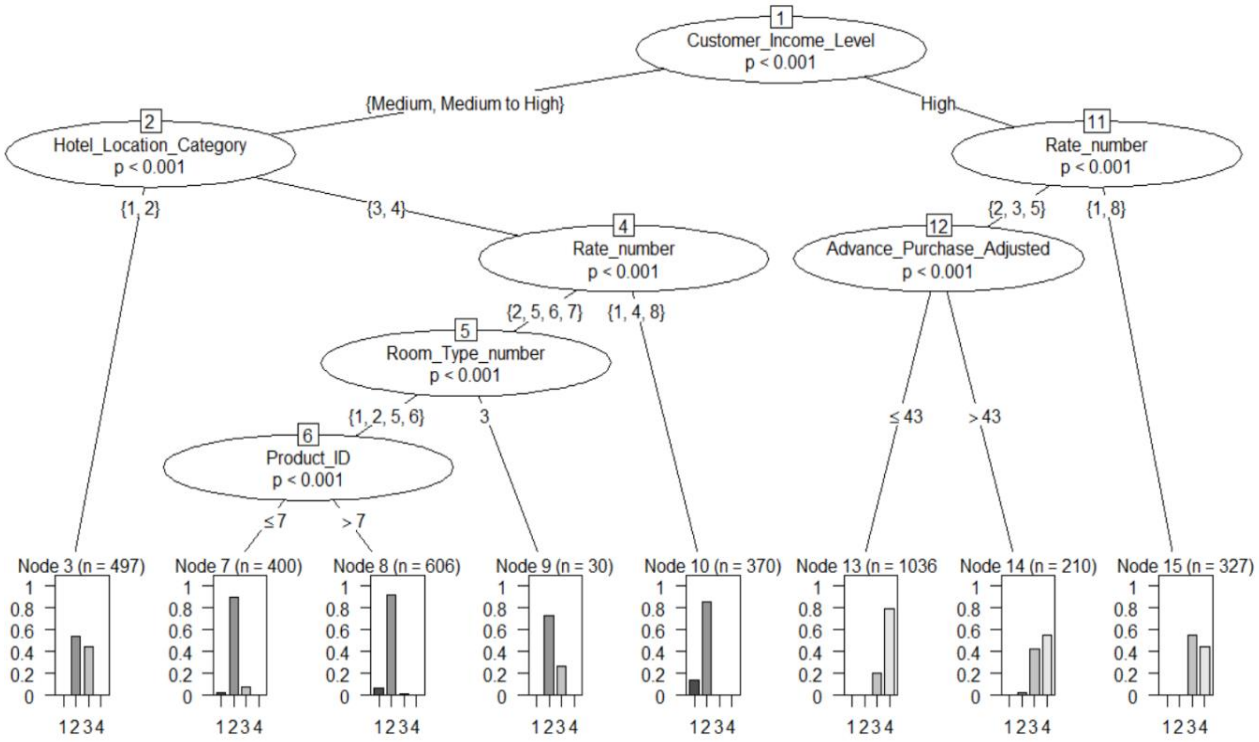
The error in the prediction model performed in neural network, calculated by the method MAE, was 43.77. It was contrasted with an error of 51.41 produced by applying linear regression. Given that the error is lower in neural network is then concluded that this machine learning method was more accurate for predictions based on the current data set that the linear regression model.

## 6.4 Decision tree

This method of prediction admits the use of both types of variables; numerical and categorical. Consequently, a selection of thirteen more representative attributes within the original data set are being considered for this analysis.

After fitting the model with the proposed variables, the resulting decision tree is as shown in figure 19. At the top of the tree, the node more relevant according to the algorithm is located. From this attribute label 1, all the other dependencies are derived in their way down through six more nodes before reaching the predicted

variable, nightly room price, depicted by single bars contained in boxes in four categorical levels.




| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 0 <= 80 | MORE THAN 80 <= 150 | MORE THAN 150 <= 300 | MORE THAN 300 |

**Figure 19. Decision tree with 15 nodes**

Each of the boxes has a probability scale ranging from 0 to 1 that indicates the likelihood for the room price category level to occur. These levels (1, 2, 3 or 4) are explained by the path followed through the branches. Consequently, the path's direction is the result of having responded to queries at the exit of each node.

Starting from the root of the tree labeled as a node 1. If customer income level is considered high, then the path follows to the right where the rate number is enquired. If the answer falls in categories 2, 3 or 5 the way goes to the left to advance purchase noted as a node 12. If the customer made the booking with less or even 43 days of anticipation, then the node 13 is reached. It shows that category 4 has about

80% of chance to occur. Therefore, is expected than customers following this path pay more than 300 dollars per room nightly.

Similarly to the procedure described in the above paragraph, all the other nodes known as leafs at the bottom of the tree, representing nightly rates, can be explained. It should also be noted that the variables; number of rooms, hotel size, party size description, length of stay adjusted and membership status, were excluded from the analysis by the algorithm for not being considered significant enough to explain the dependent variable according to the specific set of parameters under *control* function.

The prediction was executed for 870 random observations of testing data and their results were compared with the ones from the validation data set. The resulting table is showed in figure 19.

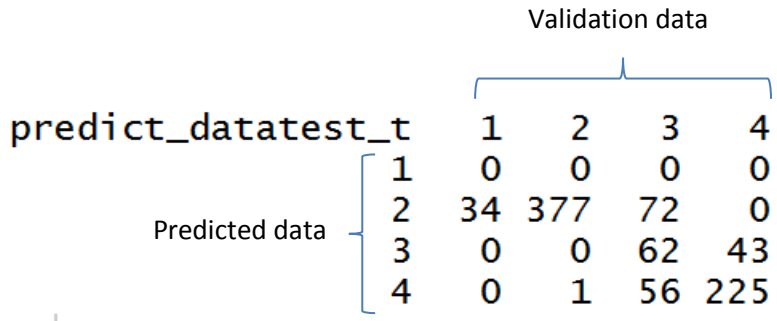



**Figure 20. Misclassification error for testing data**

Consecutive numbers 1 to 4 in the first top row as well as the first column on the left hand side of the chart on figure 20, represent four levels of nightly prices (described in figure 19). Additionally, the first row indicates validated data whereas data derived from the column on the left has been predicted by the model. The sum of all data contained in the table equals 870, a random sample of 20% that is equivalent to the testing data set as expected.

The first column of validation data counts 34 bookings which are categorized as level 1. However in the prediction, they are characterized as level 2 evidencing the existence of misclassification which basically reflects presence of errors in the model.

Based on the misclassification table is possible to state that:

From a total of 34 reservations corresponding to a level 1 or rates per room up to 80 dollars, all of them were misclassified by the model.

377 out of 378 reservations corresponding to a level 2 or rates per room between 81 up to 150 dollars were correctly classified by the algorithm.

62 out of 190 reservations corresponding to a level 3 or rates per room between 151 and 300 dollars were correctly classified.

From a total of 268 reservations corresponding to a level 4 or rates per room of 301 dollars or more, 225 were appropriately classified.

In summary, the model classified successfully 664 out of 870 reservations of the testing data set. Consequently, the error is 1- (664/870) = 23.67%.

# 7. DISCUSSION

In the context of Revenue Management - RM, deep learning techniques have emerged to enrich the quality of the process but its main focus and implementation has been on the prediction of; demand, segmentation, yields, cancellations, no shows, advance bookings and many other complex simulations.[28] However, an alternative approach like the presented in this study has not had mayor divulgation as a method for price prediction and featuring of the most relevant attributes that could explain that price.

The results of this study indicate a dependency among the pair variables analyzed according to Pearson's Chi – Square Test. In order to determine their relevance for pricing formation, a further examination was conducted and their outcomes tabulated. From this perspective, it can be stated that nightly rates in the range from 80 to 150 and more than 300 dollars represent 75% of the sales and they were mainly sold through the CRO/Hotel. This suggests that reservations department, linked directly to the hotel chain, is playing a key role in the promotion and turnover of the business.

On the other hand, more than 50% of the rooms sold through Global Distribution Channels – GDS are featured as expensive costing over 300 dollars per night. It could explain its lower contribution to the total sales (9%) and additionally a possible policy implemented from the hotel side to restrict the number of cheaper rooms and types of rate distributed through this specific channel in an attempt to diversion potential clients to its own CRO/Hotel and website.

When comparing hotel sizes, measured by their number of rooms, and distribution channel, it was found that the biggest hotels with 260 or more rooms

[28] Stanislav H. Ivanov and Vladimir Sashov, "Hotel Revenue Management – A Critical Literature Review," *ResearchGate* Vol. 60, no. 2 (2012): 187.

reported 53% of the sales and 64% of their bookings were made through the CRO/Hotel followed by their own website with 26% of the reservations.

Geographic location of the premises could also help to clarify the relationship between number of rooms and price levels. Nearly 76% of the total rooms are located in downtown areas where the big hotel size of this chain serves mainly transient or business travelers who usually are willing to pay higher rates given the convenience of having transport terminals or additional facilities nearby. In this sense, nightly room rate and number of rooms shows a positive correlation of 0.87 justifying the significant interaction between these two variables. The higher the number of rooms, the higher the nightly rate.

Amongst the eight available rates number, Rack Rate – Unrestricted Rate was related to 57% of the total bookings and sold predominantly in the range of nightly rates from 80 to 150 and 300 or more dollars. Considering that this type of rate is normally the maximum amount that the hotel charge for a room salable to any customer, it suggests that it is having a greater impact on prices intervals mentioned above at the expense of the lowest rate of up to 80 dollars.

It should be noted in this context that the type of rate for advance purchase of at least 7 days, is not relevant as most business travelers are usually booking on a short notice basis which in turn reinforces the idea of rising price levels for their reservations.

Room services category presents a value of 0% and it suggests that the hotel has restricted this kind of provision in its downtown premises which in turn would be atypical given the level of rates charged. It should also be observed that the hotel does not fall in the category of resort where stays are generally booked on longer term basis and an important part of the revenue perceived by the venue comes from trading meals, beverages and functions.

A model using neural networks was conducted in a quest for formulating a prediction model taking nightly room prices as dependent variable. The prediction shows that the values reported by the algorithm were more accurate in the first half

of the testing dataset. However, towards the end, the model failed to capture the whole dispersion displayed by the observed data. Nevertheless, the Mean Absolute Error – MEA when compared to a linear regression model is lower, 44.77 against 51.41, which means that it has more precision than the alternative method when predicting the target variable with the current dataset.

It can be mentioned that the current model can be further tuned in order to produce a more accurate forecast through the fitting of more hidden layers and nodes. Likewise, the experimentation should also include a larger amount of numerical attributes to fit the model as the current set counts only five numerical features. Additionally, software libraries like TensorFlow and Keras (compatible with RStudio) could also be implemented as they have demonstrated in similar exercises taking the prediction to a more precise level for numerical response.

In order to explain price ranges based on external and internal factors, and its selection based on importance, a classification technique known as a decision tree was utilized. The first noticeable outcome is the significant reduction in the number of attributes remaining in the final model. From a total of twelve explanatory factors fitted, only seven are displaying the price range formation. In a way, this can be an important tool for helping managers to decide in which variables they should focus on the most. On top of the chart, the customer income level appears. This is a critical factor, according to the model, hence the importance of its correct segmentation.

Another remark is the fact that the lowest price corresponding to category 1 is very unlikely to be charged. Category 2 with rates between 80 and 150 dollars, presents the highest level of probability even following multiple paths and the most expensive rates with a value of 300 dollars or more are 80% more likely to be charged to those high income customers than make a purchase on short notice. The accuracy of the model was over 76%.

An alternative method than could lead to a lower error is Random Forest. It is used for regression or classification and developed by aggregating trees. As the

model of this study, it performs a selection of attributes based on its importance and can handle a large number of features.

Two important limitations regarding this analysis have been identified. The first one refers to the shortage in the collection of data. As mentioned in chapter 4, the available datasets covers over two months of operations, condition that constrains the possibility of testing the developed models across different seasons. The second restriction lies in the fact that all five datasets were merged in one single template. By doing so, distinctive features associated with a specific type of property could have been overlooked. For example, the type of guest can influence the attributes observed in downtown venues containing large amount of rooms compared to a suburban premise where customers are linked to a different segmentation in terms of budget, length of stay or reservation channel amongst others. Another important factor that eventually limits a more precise characterization of the hotel chain is the lack of qualitative information including exact location of the premises and description of the rooms as same types might differ amongst properties.

From a practical point of view, the implantation of the above deep learning models, the analysis of correlations among attributes and software tools presented in this dissertation constitute the prosed method for price reference. It can be used on weekly or seasonal basis with the aim of helping decision makers to determine the most significant variables that in a specific point in time are explaining room prices. Although the system does not actually define an exact price for type of room given the short time frame of historical data collected, its predictions contribute to anticipate possible changes in trend rates.

Finally, it should be highlighted that at the time of writing this study, a similar analysis has not been found for comparison purposes and it could be explained by the fact that deep learning methods are still not a common practice in many small to medium size hotels and those which adopted this approach, might avoid make its methods of public domain due to confidentiality policies.

# 8. CONCLUSION

The results observed by the examination of datasets enable to state the following conclusions:

The results of Pearson's Chi – Square Test indicates dependency among the categorical variables analyzed.

The hotel chain with a total of 1220 rooms across its five properties is heavily relying on the CRO/Hotel distribution channel to boost the sailing of intermediate and high room rates in a range price from 80 to 150 and 300 dollars or more. By doing so, it avoids commissions to third parties and takes full advantage of their own work force for prompting customers with high income to purchase a premium product.

More than 50% of the rooms sold through Global Distribution Channels - GDS, corresponded to the most expensive category 4 or over 300 dollars. GDS had the lower contribution to the total sales with 9%.

It was found that the biggest hotels with 260 or more rooms reported 53% of the sales and 64% of their bookings were made through the CRO/Hotel followed by their own website with 26% of the reservations.

Rack Rate – Unrestricted Rate was related to 57% of the total bookings and sold predominantly in the range of nightly rates from 80 to 150 and 300 or more dollars.

The type of rate for advance purchase of at least 7 days is not relevant as most business travelers are usually booking on a short notice basis which in turn reinforces the idea of rising price levels for their reservations. Consequently, the Pearson coefficient reported such a slow value in the correlation of 0.16 for the association between advance purchase and nightly rates.

Nightly room rate and number of rooms shows a positive correlation of 0.87 justifying the significant interaction between these two variables. The higher the number of rooms, the higher the nightly rate.

The prediction model for nightly prices follows a clear pattern for the first half of testing observations; however its performance drops significantly as it fails to capture the wider dispersion towards the end. Nevertheless, the error when compared to a linear regression model is lower which means that it has more precision than the alternative method when predicting the target variable with the current dataset.

Decision tree allows the recognition of the most critical variable, customer income level. This variable is the result of the segmentation conducted by the hotel and becomes a key factor from which other six attributes are derived to explain the interval price formation. In average, there is a 80% chance to sell nightly rates in a range between 80 and 150 dollars to customers identified as a medium income. In contrast, the most expensive rates over 300 dollars per night are charged with a close probability of 80% to customers with high income who book in a short notice. This model presented assertiveness greater than 76% according to the testing data.

The accuracy of their results can be further improved by tuning the parameters and fitting more variables in the above supervised models. It should also be considered the implementation of additional software libraries derived from the used techniques to minimize the prediction and classification error.

# REFERENCES


A.G. Asuero, A. Sayabo and A. Gonzalez, *The Correlation Coefficient: An Overview* (Taylor and Francis Group , LLC, 2006).

Albert Bifet et al., *Machine Learning for data Streams* (London: The MIT Press, 2017)

altexsoft, "Machine Learning Redefines Revenue Management and Dynamic Pricing in Hotel Industry", altexsoft, 4 June 2018, https://www.altexsoft.com/blog/datascience/machine-learning-redefines-revenue-management-and-dynamic-pricing-in-hotel-industry/ (accessed 24 August, 2018)

Andy Thomas, "Neural Networks Tutorial - A Pathway to Deep Learning," *Adventures in Machine Learning*, March 18, 2017, http://Adventuresinmachinelearning.com/neural-networks-tutorial/. (accessed July 29, 2018).

Ben Rogojan, "How to Measure the Accuracy of Predictive Models," *acheron analytics*, May 28, 2018, http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models#. (accessed August 01, 2018).

Chaitanya Sagar, "Creating and Visualizing Neural Network in R," *Analytics Vidhya*, September 7, 2017, https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/. (accessed July 29, 2018).

David Stockburger, "Introductory Statistics: Concepts, Models and Applications," http://davidmlane.com/hyperstat/chi_square.html (accessed August 14, 2018).

Debra Adams et al., *Revenue Management: HOSPA Practitioner Series* (Bournemouth: Wentworth Jones Limited, 2013).

Foster Provost and Tom Fawcett, *Data Science for Business* (Sebastopol CA: O'Reilly, 2013), 45.

Garret van Ryzen and Kalyan Talluri, *The Theory and practice of Revenue Management* (New York: Springer, 2004).

Gary K. Vallen and Jerom J. Vallen, *Check In Check Out: Managing Hotel Operations* (Boston: Pearson, 2018), 7.

Ivanov Stanislav, *Hotel Revenue Management: From Theory to Practice*, (Varna: Zangador Ltd, 2014), 34

Mark Ferguson, Tudor Bodea, "Choice-Based Revenue Management: Data from a major Hotel Chain, *InformsPubsOnLine,* September 9, 2008, https://pubsonline.informs.org/doi/suppl/10.1287/ msom.1080.0231 (accessed May 4, 2018).

Michy Alice, "Fitting a Neural Network in R; neuralnet package," *datascience+*, September 23, 2015, https://datascienceplus.com/fitting-neural-network-in-r/. (accessed July 28, 2018).

Priya Chetti, Sudeshna, "Nominal, Ordinal and Scale in SPSS," *Project Guru,* January 16, 2015, https://www.projectguru.in/publications/nominal-ordinal-scale-spss/ (accessed August 11, 2018).

Rebecca Njeri, "What is a Decision tree Algorithm?," *Medium*, September 3, 2017, https://medium.com/@SeattleDataGuy/what-is-a-decision-tree-algorithm-4531749d2a17 (accessed August 16, 2018).

Simon Haykin, *Neural Networks and Learning Machines* (Hamilton: Pearson Prentice Hall, 2009).

SNAPSHOT TEAM, "Understanding the Basics of Hotel Revenue Management," Snapshot Travel, November 16, 2015, https://blog.snapshot.travel/understanding-the-basics-of-revenue-management (accessed June 17, 2018).

Stanislav H. Ivanov and Vladimir Sashov, "Hotel Revenue Management – A Critical Literature Review," *ResearchGate* Vol. 60, no. 2 (2012): 187

Taiwo Ayodele, *Types of Machine Learning Algorithms*, https:// www.researchgate.net/publication/ 221907660/download, (accessed August 30, 2018).

Tudor Dan Bodea, "Choice-Based Revenue Management: A Hotel Perspective" PhD diss., Georgia Institute of Technology, 2008, https://smartech.gatech.edu /bitstream/handle/1853/24739/ bodea_tudor_d_200808_phd.pdf

# Videos and tutorials sources

Amr Arafat. “R tutorial for 2-2 Examining Relationships Between Two Categorical Variables”. YouTube video, 08:09. Posted (June 2013).
https:// www.youtube.com/ watch?v=xLcrJNEYOTk

Bharatendra Rai. “Goodness of Fit and Test of Independence with R - Examples Using Chi-Square Test”. YouTube video, 13:32. Posted (April 2017).
https://www. youtube.com/watch?v=1RecjImtImY

Bharatendra Rai. “Neural Networks in R: Example with Categorical Response at Two Levels”. Youtube video, 23:06. Posted (May 2017).
https://www.youtube.com/watch?v=-Vs9Vae2KI0&t=828s

Bharatendra Rai. “Decision Tree with R | Complete Example“. YouTube video, 18:43. Posted (November 2015). https://www.youtube.com/watch?v=tU3Adlru1Ng

MarinStatsLectures. “Chi-Square Test, Fishers Exact Test, and Cross Tabulations in R (R Tutorial 4.7)”. YouTube video, 03:54. Posted (August 2013).
https://www. youtube.com / watch?v=POiHEJqmiCo

Math Meeting. “Chi Square Test - with contingency table “. YouTube video, 17:03. Posted (May 2015). https://www.youtube.com/watch?v=misMgRRV3jQ

# APPENDIX A. RStudio Coding Transcripts

## Correlations

```
hotel<- read.csv(file.choose(), header = T, sep=";")
head(hotel)


hotel_filtered <- hotel[ , c(3,4,5,6,7,9,13,14,15,16,18,19,20,21,23,25,27,28,29)]
str(hotel_filtered)
head(hotel_filtered)

hotel_filtered$Hotel_Location_Category<- as.factor(hotel_filtered$Hotel_Location_Category)
hotel_filtered$Hotel_Size_Category<- as.factor(hotel_filtered$Hotel_Size_Category)
hotel_filtered$Income_Level_Category<- as.factor(hotel_filtered$Income_Level_Category)
hotel_filtered$Distribution_Channel_number<-
as.factor(hotel_filtered$Distribution_Channel_number)
hotel_filtered$Party_Size_Category<- as.factor(hotel_filtered$Party_Size_Category)
hotel_filtered$Nightly_Rate_Category<- as.factor(hotel_filtered$Nightly_Rate_Category)
hotel_filtered$Rate_number<- as.factor(hotel_filtered$Rate_number)
hotel_filtered$Room_Type_number<- as.factor(hotel_filtered$Room_Type_number)
hotel_filtered$Merge_Indicator<- as.factor(hotel_filtered$Merge_Indicator)
hotel_filtered$Membership_Status<- as.factor(hotel_filtered$Membership_Status)
hotel_filtered$VIP_Membership_Status<- as.factor(hotel_filtered$VIP_Membership_Status)


      #######  Nigtly rate Category Vs. Distribution Channel  ######


## Nightly Rate category
T_Rate<-table(hotel_filtered$Nightly_Rate_Category)

## frequencey table for T_Rate

frq_T_Rate<- T_Rate/4346
barplot(frq_T_Rate, beside=t, col=rainbow(4),
            main= "Room Rate Category",
            ylab= "Frecuency",
            xlab= "Price range")

legend("topright",c("0 <= 80",
              "MORE THAN 80 <= 150",
              "MORE THAN 150 <= 300",
              "MORE THAN 300"),
              pch =15, col = rainbow(4))


## combining frequency and table Rate

cbind(T_Rate, frq_T_Rate)
```

```
##    Distribution channel
T_channel<-table(hotel_filtered$Distribution_Channel)
frq_Channel<- T_channel/4346
barplot(frq_Channel, beside=t, col=rainbow(3),
              main= "Distribution Channel",
              ylab= "Frecuency",
              xlab= "Distribution Channel")

cbind(T_channel, frq_Channel)


## Distribution. Contingency table

test_table<- table(hotel_filtered$Nightly_Rate_Category,
                   hotel_filtered$Distribution_Channel)

ftest_table<- test_table/4346
round(ftest_table,2)

barplot(ftest_table, beside = T, col = rainbow(4),
                            ylab = "Frecuency",
                            xlab = "Distribution Channel")

legend("topright",c("0 <= 80",
                    "MORE THAN 80 <= 150",
                    "MORE THAN 150 <= 300",
                    "MORE THAN 300"),
                    pch =15, col = rainbow(4))


## Conducting Chi-squared test with significant level of 0.05

# Ho. Nightly rate and distribution channel are independent
# Ha. Nightly rate and distribution channel are dependent

chisq.test(test_table)



#######  Type of hotel size Vs. Distribution Channel  ######


## Type of Hotel size
T_Hotel<-table(hotel_filtered$Hotel_Size)
frq_T_Hotel<- T_Hotel/4346
barplot(frq_T_Hotel, beside=t,col=rainbow(3),
                    main= "Distribution of bookings by hotel size",
                    ylab= "Frecuency",
                    xlab= "Hotel size")

legend("topright",c("Big - 260 or more rooms",
                    "Medium - 71 to 259 rooms",
                    "Small - up to 70 rooms"),
                    pch =15, col = rainbow(3))

cbind(T_Hotel, frq_T_Hotel)
```

```
##    Distribution channel
T_channel<-table(hotel_filtered$Distribution_Channel)
frq_Channel<- T_channel/4346
barplot(frq_Channel, beside=t, col=rainbow(3),
                     main= "Distribution Channel",
                     ylab= "Frecuency",
                     xlab= "Distribution Channel")

cbind(T_channel, frq_Channel)


## Distribution. Contingency table

DC_HS<- table(hotel_filtered$Distribution_Channel,
              hotel_filtered$Hotel_Size)

DC_HS_table<- DC_HS/4346
round(DC_HS_table,2)

barplot(DC_HS_table, beside = T, col = rainbow(3),
                       main= "Distribution Channel",
                      ylab = "Frecuency",
                      xlab = "Hotel Size")

legend("topright",c("CRO/Hotel","GDS","WEB"),
               pch =15, col = rainbow(3))



## Conducting Chi-squared test with significant level of 0.05

# Ho. Hotel size and distribution channel are independent
# Ha. Hotel size and distribution channel are dependent


chisq.test(DC_HS)



#######  Nightly rate Category Vs. Rate number (Type of booking)  ######

## Nightly Rate category
T_Rate<-table(hotel_filtered$Nightly_Rate_Category)

## frequencey table for T_Rate
frq_T_Rate<- T_Rate/4346
barplot(frq_T_Rate, beside=t,col = rainbow(4),
          main= "Distribution of bookings by Rate category",
          ylab= "Frecuency",
          xlab= "Rate category")

legend("topright",c("0 <= 80",
                "MORE THAN 80 <= 150",
                "MORE THAN 150 <= 300",
                "MORE THAN 300"),
                pch =15, col = rainbow(4))
```

```
# more than 40% of bookings fell in the price category 2
## combining frequency and table Rate
cbind(T_Rate, frq_T_Rate)

##    Rate Number
R_Number<-table(hotel_filtered$Rate_number)
frq_R_Number<- R_Number/4346
barplot(frq_R_Number, beside=t,
          col = rainbow(8),
          main= "Distribution of bookings by Rate number",
          ylab= "Frecuency",
          xlab= "Rate number")

legend("topright",c("Adv. Purchase at least 7 day. Fully Restricted Rate",
                "Rack Rate - Unrestricted Rate",
                "Rack Rate with Add. Hotel Services",
                "Discount Rate Less Restricted Than Adv. Purchase",
                "Accom. with Frequent Traveler Rewards",
                "Accom. with In Room Services",
                "Accom.Airport/Airline Services",
                "Accom. with City/Weekend Activities"),
                pch =15, col = rainbow(8))

cbind(R_Number, frq_R_Number)


## Distribution. Contingency table

DR_Number<- table(hotel_filtered$Rate_number,
               hotel_filtered$Nightly_Rate_Category)

DR_Number_table<- DR_Number/4346
round(DR_Number_table,2)

barplot(DR_Number_table, beside = T, col = rainbow(8),
                    ylab = "Frecuency",
                    xlab = "Rate Category")

legend("bottomleft", c("Adv. Purchase at least 7 day. Fully Restricted Rate",
                "Rack Rate - Unrestricted Rate",
                "Rack Rate with Add. Hotel Services",
                "Discount Rate Less Restricted Than Adv. Purchase",
                "Accom. with Frequent Traveler Rewards",
                "Accom. with In Room Services",
                "Accom.Airport/Airline Services",
                "Accom. with City/Weekend Activities"),
                pch =15, col = rainbow(8))


## Conducting Chi-squared test with significant level of 0.05

# Ho. Nightly rate category and type of rate are independent
# Ha. Nightly rate category and type of rate are dependent
chisq.test(DR_Number)
```

```
### Correlation with Pearson coefficient for numerical variables   ###


n_variable <- hotel_filtered[, c(2,9,10,12,14)]

str(n_variable)

hotel_cor <- cor(n_variable, method = "pearson")
round(hotel_cor,2)
```

## Neural network Coding

```
## Importing initial file into R ##

hotel<- read.csv(file.choose(), header = T, sep = ";")

hotel_int <- hotel[, c(4,7,15,16,19,21)]
head(hotel_int)
str(hotel_int)

## Random Sampling ##

samplesize = 0.80 * nrow(hotel_int)
set.seed(80)
index<-sample(seq_len(nrow(hotel_int)), size = samplesize)


# Create training and test set

datatrain = hotel_int[ index, ]
datatest = hotel_int[ -index, ]


## Scale data for neural network

max = apply(hotel_int , 2 , max)
min = apply(hotel_int , 2 , min)

scaled = as.data.frame(scale(hotel_int, center = min, scale = max - min))

head(scaled)


# creating training and test set

trainNN = scaled[index , ]
testNN = scaled[-index , ]


# fit neural network

library(neuralnet)
```

```
set.seed(2)

NN <- neuralnet(trainNN$Nightly_Rate~ trainNN$Number_Of_Rooms +
                        trainNN$Product_ID +
                        trainNN$Advance_Purchase_Adjusted +
                        trainNN$Party_Size +
                        trainNN$Length_of_Stay_Adjusted,
                        data = trainNN,
                        hidden = c(2,1),
                        act.fct = "logistic",
                        linear.output = TRUE)

# plot neural network

plot(NN)

## Prediction using neural network

predict_testNN = compute(NN, testNN[,c(-6)])

# Scaling back to standard values to make a meaningful comparison #

predict_testNN = (predict_testNN$net.result * (max(hotel_int$Nightly_Rate)-
                                    min(hotel_int$Nightly_Rate)))+
                                    min(hotel_int$Nightly_Rate)

# Prediction Vs. testing data for quality visualization #

plotNN <- plot(datatest$Nightly_Rate, predict_testNN, col='orange',
                                  pch=16,
                                  cex.axis = 1.0,
                                  main = "Prediction Quality vizualization on Testing Data",
                                  ylab = "predicted Nightly Rate",
                                  xlab = "Testing Nightly Rate",
                                  abline(0,1))

plot(datatest$Nightly_Rate, col='blue',
    pch=16,
    cex.axis = 1.0,
    main = "Validated (Observed) Nightly Rate",
    ylab = "Nightly Rate")

plot(predict_testNN, col='red',
    pch=16,
    cex.axis = 1.0,
    main = "Predicted Nigtlhy Rate",
    ylab = "Nightly Rate")

# Quality by error

err_predict_NN = mean(abs(predict_testNN - datatest$Nightly_Rate))

#  = 43.77  #

## Prediction Quality comparison with Linear Regression ##
```

```
TRN_LM = scaled[index , ]
TST_LM = scaled[-index , ]


model = lm(Nightly_Rate ~ Number_Of_Rooms+
                    Product_ID +
                    Advance_Purchase_Adjusted +
                    Party_Size +
                    Length_of_Stay_Adjusted,
                    data = TRN_LM)

predict_testLM = predict(model, TST_LM[,c(-6)])

predict_testLM = (predict_testLM * (max(hotel_int$Nightly_Rate) -
                            min(hotel_int$Nightly_Rate)))+
                            min(hotel_int$Nightly_Rate)

err_predict_LM = mean(abs(predict_testLM - datatest$Nightly_Rate))

##  = 51.41  ##

err_base = mean(abs(mean(datatrain$Nightly_Rate) - datatest$Nightly_Rate))

# = 119.67  #
```

## Decision Tree

```
## Creating index variable

hotel <- read.csv(file.choose(), header=T, sep=";")
hotel_tree <- hotel[ , c(3,4,5,7,8,14,15,17,19,20,23,25,28)]
str(hotel_tree)
head(hotel_tree)

## Converting variables into the right classification

hotel_tree$Hotel_Location_Category<-as.factor(hotel_tree$Hotel_Location_Category)
hotel_tree$Nightly_Rate_Category<-as.factor(hotel_tree$Nightly_Rate_Category)
hotel_tree$Rate_number<-as.factor(hotel_tree$Rate_number)
hotel_tree$Room_Type_number<-as.factor(hotel_tree$Room_Type_number)
hotel_tree$Membership_Status<-as.factor(hotel_tree$Membership_Status)

# Decision tree with Party
install.packages("party")
library(party)

# Random sampling

samplesize = 0.80 * nrow(hotel_tree)
set.seed(80)
index_tree <- sample(seq_len(nrow(hotel_tree)), size = samplesize)
```

```
# Create training and test datasets

train_hotel = hotel_tree[index_tree, ]
test_hotel = hotel_tree[-index_tree, ]


# fit decision tree

h_tree <- ctree(Nightly_Rate_Category~ Hotel_Location_Category+
                Number_Of_Rooms+
                Hotel_Size+
                Product_ID+
                Customer_Income_Level+
                Distribution_Channel+
                Advance_Purchase_Adjusted+
                Party_Size_Description+
                Length_of_Stay_Adjusted+
                Rate_number+
                Room_Type_number+
                Membership_Status,
                data = train_hotel,
                controls = ctree_control(mincriterion = 0.99, minsplit = 1000))

h_tree

plot(h_tree, cex = .4)


## Prediction on testing data ##

predict_datatest_t <- predict(h_tree, test_hotel)

plot(predict_datatest_t)

## Comparison between testing and predictive data ##

head(predict_datatest_t)

head(test_hotel$Nightly_Rate_Category)

plot(predict_datatest_t,test_hotel$Nightly_Rate_Category)


   ## Misclassification error for train data ##

table_hotel<- table(predict(h_tree),train_hotel$Nightly_Rate_Category)
table_hotel
1-sum(diag(table_hotel))/sum(table_hotel)
# = 0.2399 #
# The misclassification error is about 24.36% based on the training data
## Misclassification error for testing data ##
table_hotel2<- table(predict_datatest_t,test_hotel$Nightly_Rate_Category)
table_hotel2
1-sum(diag(table_hotel2))/sum(table_hotel2)
# = 0.2367 #
```