



**UTM**  
UNIVERSITI TEKNOLOGI MALAYSIA

**University of Technology, Malaysia  
Faculty of Engineering  
School of Computing  
Semester 2019-2020 2**

SECI2143-02  
Probability and Statistical Data Analysis  
Project 2- Report

**Statistical Data Analysis on Hotel Average Daily Rate and  
Bookings**

**Lecturer:**  
Dr. Chan Weng Howe

**Name:**  
Muhammad Irfan Daniel Bin Abd Karim (A19EC0197)

# Table of Contents

1.0 Introduction .....	3
2.0 Hypothesis Testing .....	4
2.1 Hypothesis Testing Using One Sample .....	4
2.2 Correlation Analysis .....	7
2.3 Regression Analysis .....	10
2.4 Chi Square Test of Independence .....	13
3.0 Discussion .....	18
4.0 Conclusion .....	19
5.0 References .....	20

# 1.0 Introduction

In this Project, I took a datasets from Kaggle.com about the Hotel Booking Demands at some regions in Portugal. The secondary data set was analyzed and arranged by one of Kaggle's user, Jesse Mostipak. The data set was collected by a group of expertise (updated on 5 October 2018). The data was acquired from the Extractions of hotel's Property Management System (PMS)- SQL databases. The data sources location was from Resorts Hotel at the region of Algarve and City Hotel at the city of Lisbon. The person who gathered this data were:

1. Nuno Antonio (A, B)
2. Ana de Almedia (A, C, D)
3. Luis Nunes (A, B, D)

*(Note: the alphabet is the key of the universities)*

All of them have different and many universities background such as:

- A. Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
- B. Instituto de Telecomunicações, Lisbon, Portugal
- C. CISUC, Coimbra, Portugal
- D. ISTAR-IUL, Lisbon, Portugal

Basically, This data article describes two datasets containing data on hotel demand. One of the hotels (H1) is a Resort Hotel, and the other (H2) is a City Hotel. Both datasets share the same structure, with 31 variables describing H1 with 40060 observations and H2 with 79330. Every observation represents a hotel reservation. Both datasets include bookings scheduled to arrive between 1 July 2015 and 31 August 2017, including bookings that have arrived successfully, and reservations that have been cancelled. Since this is real data from the hotel, all data elements relating to hotel or costumer identification have been removed. These datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields due to the scarcity of real business data for scientific and educational purposes.

The main purposes of this Project is to predict the requirements in order to increase the Hotel Demands in future for business purposes. As an example when to make the promotions and how we can set the hotel room price according to time and date. The hotel owner can predict the demand of the hotel room in future. I will find the correlation and regression of lead time of hotel booking or average daily rate with the demand of the hotel

room. The mean and variance acquired will be the best indicator to find the best time to book the hotel from the proportion of the hotel (Resort Hotel and City Hotel).

## 2.0 Hypothesis Testing

Based on the selected variables from the Project Proposal, these are the proposed analysis that I planning to do:

1. Hypothesis Testing Using One Sample
2. Correlation Analysis
3. Regression Analysis
4. Chi Square test of Independence

The chosen variables were:

1. Hotel bookings (ratio)
2. Arrival date by Year (interval)
3. Lead time in days (ratio)
4. Average Daily Rate (ratio)

### 2.1 Hypothesis Testing Using One Sample

In statistics, a hypothesis is a claim or statement about a property of a population. A hypothesis test (or test of significance) is a standard procedure for testing a claim about a property of a population. To conduct a Hypothesis Testing, we must establish:

1. Hypothesis statement
2. Test Statistic
3. Significance Level/ Level of confidence
4. Conclusion

In this datasets, I will test the sample mean of the hotel Average Daily Rate (ADR) in Portugal based on the population mean I got from the websites. As my datasets is from year 2015-2017, so the population mean for ADR on that years is €101 . I want to check whether this sample has a mean greater than €101 by using 0.05 significance level (if they have the mean greater than €101 then the ADR is quite expensive).

**Hypothesis:**

$H_0: \mu = €101$

$H_1: \mu > €101$

**Test statistics:**

$\alpha = 0.05$

Critical value=  $z_{0.05} = 1.644854$

$z = 5.682631$

**R console:**

```
> #2.1 Hypothesis on one sample-mean
```

```
> #start
```

```
> adr<-hotel_bookings$adr
```

```
> mean(adr)
```

```
[1] 101.8311
```

```
> sd(adr)
```

```
[1] 50.53579
```

```
> nrow(hotel_bookings)
```

```
[1] 119390
```

```
> alpha<-0.05
```

```
> x<-mean(adr)
```

```
> o<-sd(adr)
```

```
> n<-nrow(hotel_bookings)
```

```
> z<-(x-101)/(o/sqrt(n))
```

```
> z.alpha=qnorm(1-alpha)
```

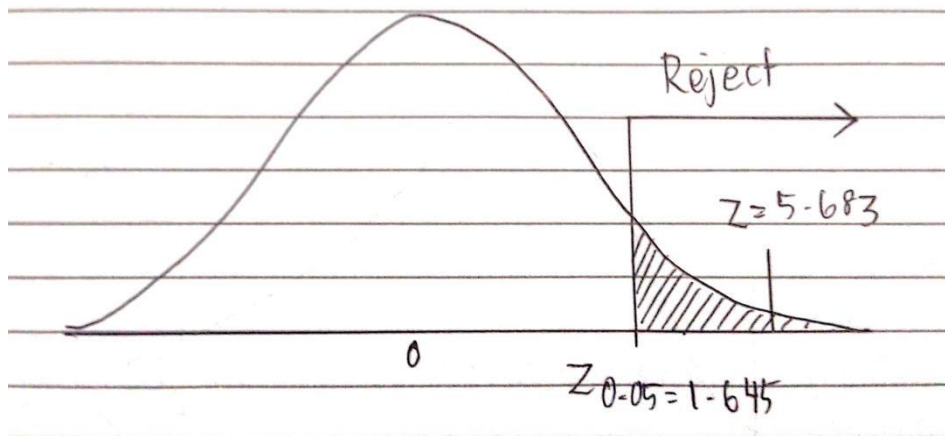
```
> z
```

```
[1] 5.682631
```

```
> z.alpha
```

```
[1] 1.644854
```

**Graph:**



**Result:**

Reject  $H_0$ .

**Conclusion:**

Statistic value,  $z = 5.683 > z_{0.05} = 1.645$ . The value falls within the rejection region. There is sufficient evidence to conclude that  $\mu > \text{€}101$ . These results suggest that the ADR in this hotel sample is more expensive than the population mean at 0.05 significance level.

## 2.2 Correlation Analysis

Correlation is a measure of the statistical relationship between two comparable variables or quantities. When two sets of data are strongly linked together, they have a high correlation. Scatter plot is used to show the relationship between two variables. In this datasets, there are two variables that I thought that they have relation to each other which is ADR and the lead time of the booking (days). This is because in the real life situation, the earlier the time of the booking, the cheaper the ADR that we can get so I want to know if this statement can be accepted or not with 0.05 significance level.

**Hypothesis:**

$H_0: \rho = 0$  (no linear correlation between lead time and ADR)

H1:  $p \neq 0$  (linear correlation exists)

**Test statistic:**

$\alpha = 0.05$

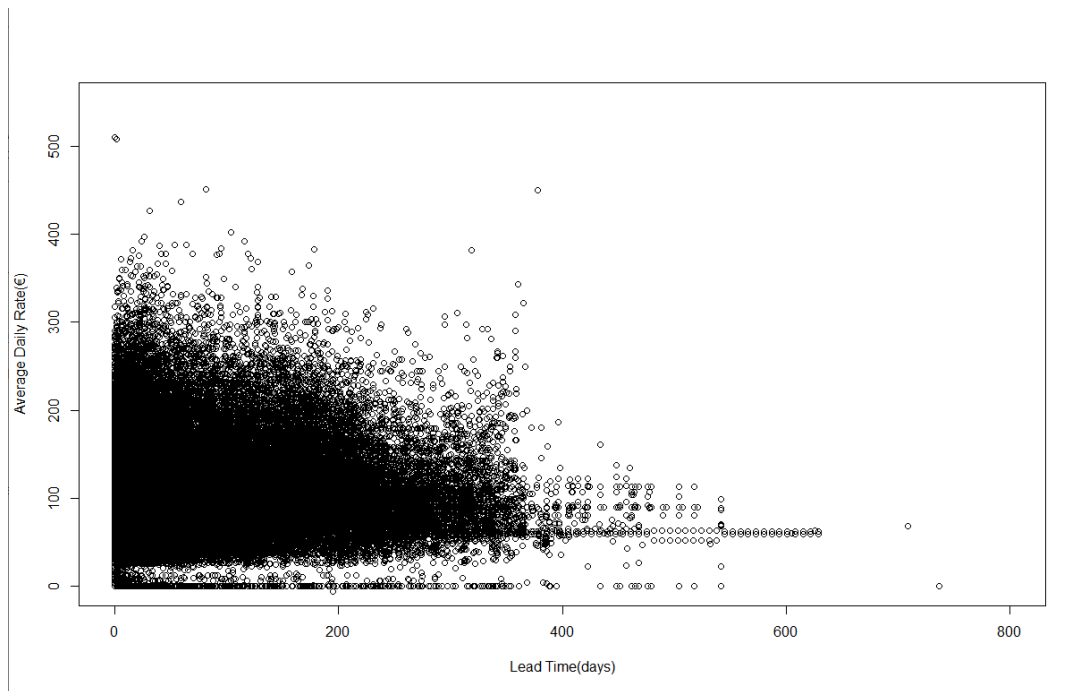
$t_{0.025, 119388} = \pm 1.960$

$t = -21.83816$

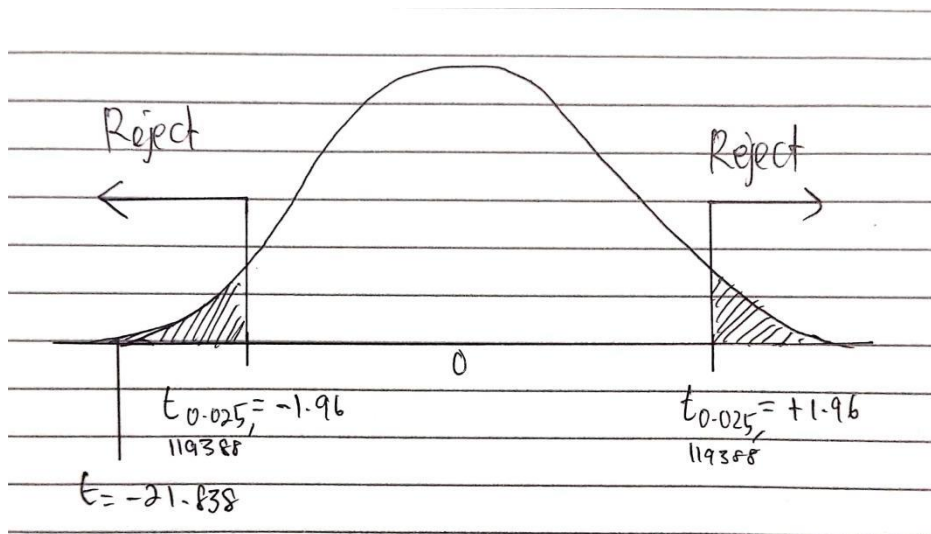
**R Console:**

```
> #2.2 Correlation
> leadtime<-hotel_bookings$lead_time
> cor(leadtime, adr)
[1] -0.06307685
> plot(leadtime, adr, xlim = c(0,800),
+      ylim = c(0, 550), xlab = "Lead Time(days)",
+      ylab = "Average Daily Rate(€)")
> r<-cor(leadtime, adr)
> t<-r/(sqrt((1-r^2)/(n-2)))
> t
[1] -21.83816
```

**Scatter Plot:**



**Graph:**



**Result:**

Reject  $H_0$ .

Sample correlation coefficient,  $r = -0.06307$

**Conclusion:**

I can reject the  $H_0$  because the  $t$  of test statistic  $= -21.83816 < t_{0.025, 119388} = -1.960$ . the test statistic result falls within the critical region. Thus, there is sufficient evidence of a linear relationship between the lead time and the Average Daily Rate at 0.05 significance level.

From the Scatter plot, we can see that both variables has a negative correlation and weak linear relationship with  $r = -0.06307$ . This means that the dependent variable, Average Daily Rate is not too influenced by the independent variable, Lead time. The scatter plot become so dark because I use a large sample of hotels which is  $n = 119390$ .

## 2.3 Regression Analysis

Regression is a little bit different than correlation. Regression analysis is uses to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the independent variable

For this data sets, I use the correlation variables as it is easier to do the regression because we already have the dependent and independent variables. This is a simple regression because I only use single independent variables which is Lead Time. Again, does the relationship between ADR and Lead Time appear to be linear?

**R Console:**

```
> #2.3 Regression
```

```
> model<-lm(adr~leadtime)
```

```
> model
```

Call:

```
lm(formula = adr ~ leadtime)
```

Coefficients:

(Intercept)	leadtime
104.93370	-0.02983

```
> plot(leadtime, adr, xlim=c(0,750),
```

```
+      ylim=c(0,550))
```

```
> abline(model)
```

```
> summary(model)
```

Call:

```
lm(formula = adr ~ leadtime)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-105.5	-31.4	-7.2	23.9	5296.1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	104.933697	0.203692	515.16	<2e-16 ***
leadtime	-0.029829	0.001366	-21.84	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.44 on 119388 degrees of freedom

Multiple R-squared: 0.003979, Adjusted R-squared: 0.00397

F-statistic: 476.9 on 1 and 119388 DF, p-value: < 2.2e-16

#### **Linear Regression Model:**

$$\hat{y} = 104.93370 - 0.02983 x$$

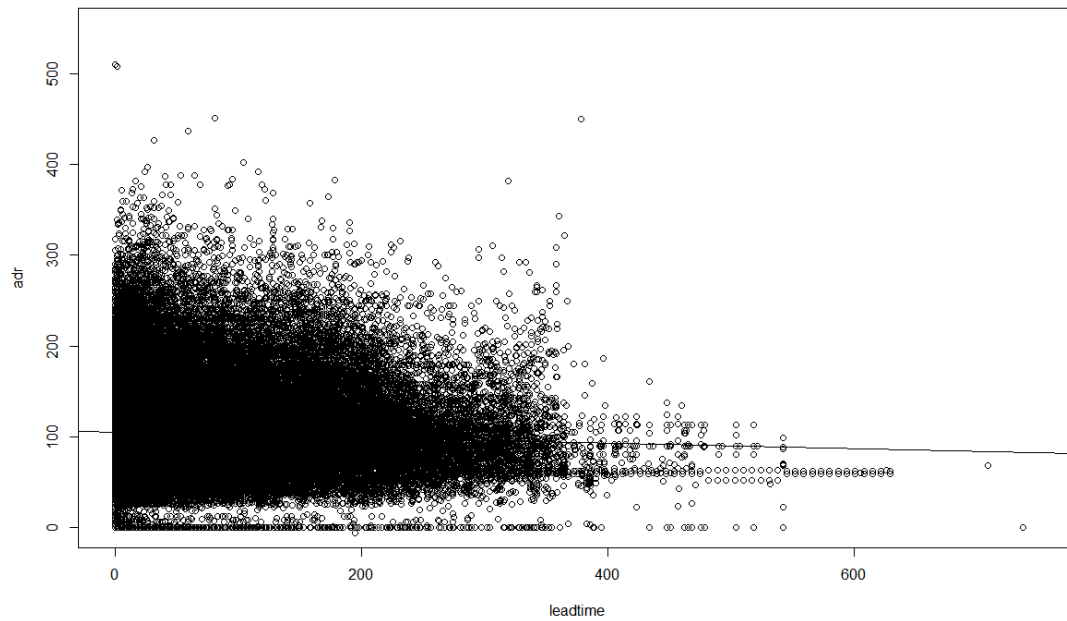
The intersection coefficient,  $b_0 = 104.93370$  is the estimated value of ADR when the value of Lead Time is zero. Here there is a lot of ADR when the lead time is 0 but we chose the most concentrated plot to be the  $b_0$ . Thus the value €104.93 is not necessary the price or ADR if the lead time is 0.

The slope coefficient,  $b_1 = -0.02983$  measures the estimated change in the average value of ADR as a result of a one-unit change in Lead time. Here, the value of ADR decrease by €0.02983, on average, for each additional of one days of Lead Time.

Coefficient of Determination,  $R^2 = 0.00397$ .

Only 0.397% of the variation in ADR for hotel is explained by the variation in the Lead Time. This coefficient of determination gives a very weak relationship between the variables.

#### **Scatter plot and Regression line:**



The regression line shows that they have a weak and negative linear relationship.

## 2.4 Chi Square Test of Independence

### One Way Contingency Table

**Hypothesis:**

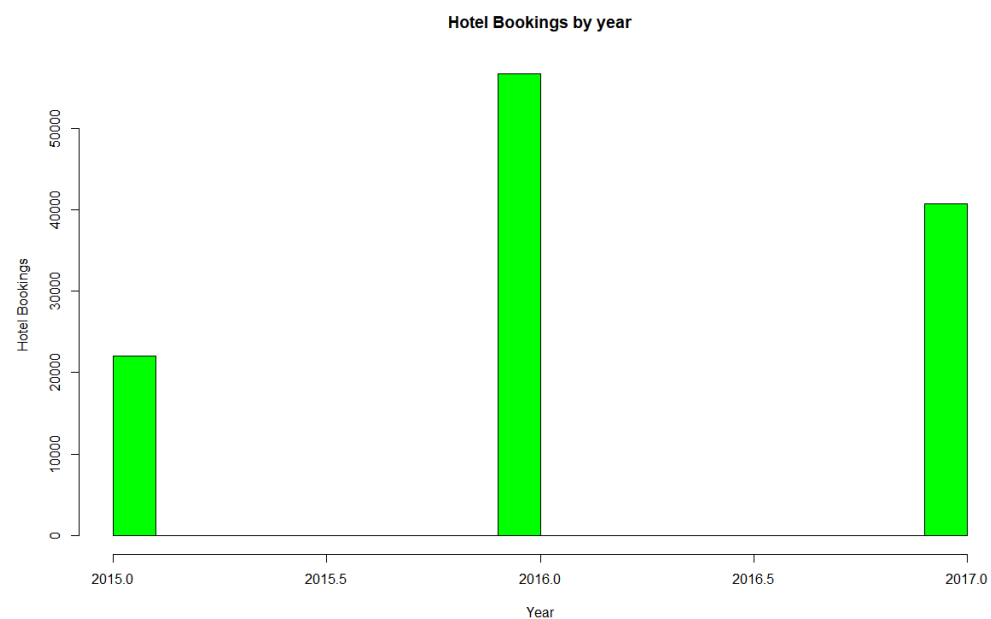
The hotel bookings occur with the same proportion at  $\alpha=0.05$ . Therefore:

$H_0: p_1=p_2=p_3$

$H_1$ : At least 1 of the 3 proportion is different from each others.

Years	2015	2016	2017	Total
No. of bookings, $P(i)$	21996	56707	40687	119390

*(The number of booking got from R Console)*



#### Calculations:

Expected frequency,  $E = 39796.66667$

$\alpha = 0.05$

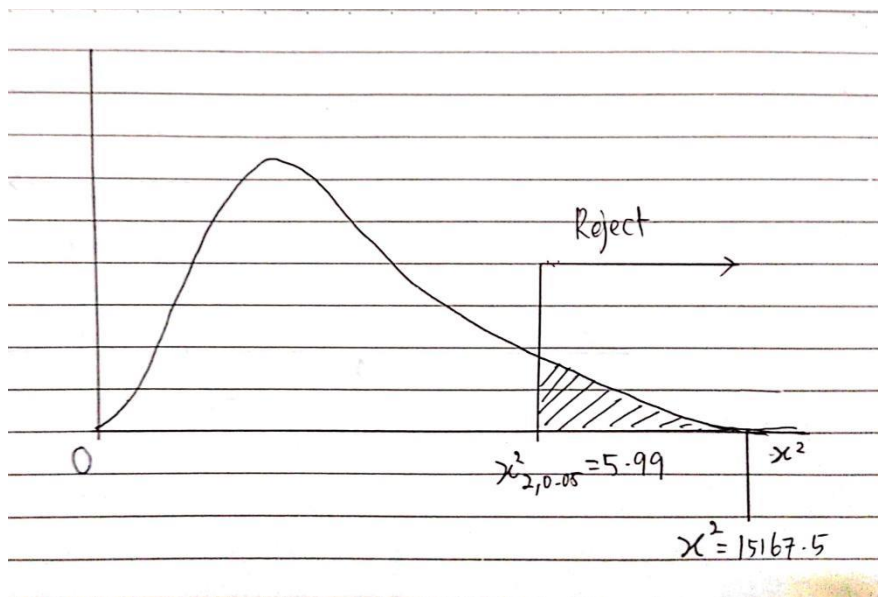
$k-1 = 3-1 = 2$

$\chi^2_{2, 0.05} = 5.991465$

#### Test Statistics:

$\chi^2 = 15167.5$

**Graph:**



**Results:**

Reject  $H_0$

**Conclusion:**

Test statistic,  $\chi^2 = 15167.5 > \chi^2_{2, 0.05} = 5.991465$ . It falls within the critical region. That is, we reject the claim that the hotel bookings occur with the equal proportions (frequency) on the 3 years.

**R Console:**

```
> #2.4 Chi Square Test  
> year<-hotel_bookings$arrival_date_year  
> hist(year, col="green",  
+       xlim=c(2015, 2017), xlab="Year",  
+       ylab="Hotel Bookings",  
+       main = "Hotel Bookings by year")
```

```

> bins<-seq(2014.5,2017.5, by=1)
> years<-cut(year, bins)
> table(years)
years
(2014.5,2015.5] (2015.5,2016.5] (2016.5,2017.5]
                21996           56707           40687
> noYear<-c(21996, 56707, 40687)
> output<-chisq.test(noYear, correct = FALSE)
> output

```

Chi-squared test for given probabilities

```

data:  noYear
X-squared = 15167, df = 2, p-value < 2.2e-16

```

```

> expprob<-sum(noYear)/3
> expyear<-c(expprob, expprob, expprob)
> exp<-((noYear-expyear)^2)/expyear
> x2<-sum(exp)
> alpha<-0.05
> x2.alpha<-qchisq(alpha, df=2,
+                   lower.tail = FALSE)
> x2.pvalue<-pchisq(x2, df=2,
+                   lower.tail = FALSE)
> x2
[1] 15167.5
> x2.alpha
[1] 5.991465

```

## 3.0 Discussion

To calculate the number of hotel bookings, I calculate the number of rows in the .csv file using R. My samples is quite large which is 119390 to make it more similar with the population. During my coding in R, I realized that to make a date as variable is not an easy task so I changed my outcome and Proposed analysis in the Project Proposal. The Proposed analysis that I changed is from Goodness-of-Fit Test to Chi Square Test (One Way Contingency Table). My new outcomes in this report are:

1. To test whether the mean of ADR in this hotel sample is higher than the hotel population in Portugal
2. Able to find the correlation and regression between Lead Time and the ADR
3. Figure out on the frequencies of hotel bookings through out three years (2015-2017)

From the potential variables in the Project Proposal, I only chose 4 variables. The other variables is not suitable for the test that I proposed. Most of them is not ratio or interval variable. It can cause many difficulties and confusion in R. When looking at my scatter plot, there is a lot of circle that made it dark at a certain place. This is also because of the large sample from the .csv file. The way I create the hypothesis in 2.1 is from websites. I searched for the population mean for hotel ADR in Portugal. For 2.4, I created the hypothesis by myself.

## 4.0 Conclusion

From my inferential statistics in this Project 2, I can conclude that:

1. The mean ADR (Euro) in this hotel sample is more expensive than the population hotel in Portugal.
2. There is a weak and negative linear relationship between the Lead Time (days) with ADR with the Sample correlation coefficient,  $r = -0.06307$ . So in my hotel samples location which is Algarve and Lisbon during 2015 to 2017, the Hotel Management did not decrease the ADR if we book the hotel earlier.
3. For regression analysis, I got equation  $\hat{y} = 104.93370 - 0.02983 x$  for the linear regression model. The regression gives very weak Coefficient of Determination,  $R^2$  which is 0.00397.

4. Only 0.397% of the variation in ADR for hotel is explained by the variation in the Lead Time.
5. The number of hotel booking does not occur in same frequencies from 2015 to 2017.

## 5.0 References

1. Jennifer Luty, (27 March 2020), Average daily hotel rate in Lisbon from 2011 to 2019 (in euros).

Retrieved from <https://www.statista.com/statistics/545281/daily-hotel-rates-lisbon/>

2. Jesse Mostipak, (February 2020), Hotel booking demand.

Retrieved from <https://www.kaggle.com/jessemostipak/hotel-booking-demand>

3. Nuno Antonio, Ana De Almeida, Luis Nunes, (February 2019), Hotel booking demand datasets.

Retrieved from <https://www.sciencedirect.com/science/article/pii/S2352340918315191>

4. Chan Weng Howe, (2020), R tutorial pdf (Part 1 to Part 5).