DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2016

Factors Affecting the Price of Airline Tickets

A case study within regression analysis

IDA JANÉR

STINA KARLSSON

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Factors Affecting the Price of Airline Tickets

A case study within regression analysis

I D A J A N É R
S T I N A K A R L S S O N

Degree Project in Applied Mathematics and Industrial Economics (15 credits)
Degree Progr. in Industrial Engineering and Management (300 credits)
Royal Institute of Technology year 2016
Supervisors at KTH: Thomas Önskog, Jonatan Freilich
Examiner: Henrik Hult

TRITA-MAT-K 2016:18
ISRN-KTH/MAT/K--16/18--SE

Royal Institute of Technology
SCI School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Faktorer som påverkar flygbiljettspriser

Sammanfattning

Denna rapport undersöker vilka faktorer som har en påverkan på flygbiljettspriser och vad deras påverkan är. Detta görs genom att utföra en multipel linjär regressionsanalys. Bakgrunden till detta är att kunna fastställa när det är bäst att boka och resa, för att få det lägsta priset. Olika regressioner genomfördes per destination, där resultaten för Bangkok och London specifikt presenterades, av de sex destinationer som undersöktes. Tre av dessa är inom Europa och tre är på ett längre avstånd. Baserat på dessa resultat undersöktes konceptet i att hitta två motsvarande generella modeller, vilket visade sig vara möjligt, till en viss nivå. Det visade sig också att samma faktorer hade en inverkan på priset för alla destinationer, dock varierade deras inverkan något. Dessa faktorer var veckodagen för ut- och hemresa, avgångsflygplatsen, avresemånaden, antalet dagar mellan bokning och utresedatumet, och slutligen, om det var ett lov under resan.

Abstract

This report investigates which factors have an impact on the price of airline tickets and what their impact is. This is done by performing a multiple linear regression analysis. The logic is to be able to establish when it is the best time to book and to leave, as to attain the lowest price. Regressions were run per destination, where the results for Bangkok and London were specifically presented, out of six destinations examined. Three of these are within Europe and three are at a longer distance. Based on the results, the concept of finding two corresponding general models was also explored and it was found that this was possible, to some extent. The same factors were found to have an impact on all destinations, although their impacts varied some. These factors were the weekday of the outbound and inbound flight, the departure airport, the month of the outbound flight, the number of days between the booking and the outbound flight, and finally, if there was a holiday during the trip.

Preface

We, Stina Karlsson and Ida Janér, would like to thank our two supervisors for this study and report, from The Royal Institute of Technology. Thomas Önskog, at the Institution for Mathematical Statistics, has provided us with his insights on regression analysis, and similarly, Jonatan Freilich, at the Institute for Industrial Management, has given us his feedback on the marketing part of this report.

We would also like to thank Mattias Nyman, the CEO of Flygresor.se, who has provided his insights into the airline industry, and his colleague Fredrik Ekedahl, who has been helpful in terms of data processing.

Finally, we would like to thank Tomas Svensson, who has helped us handling the large sets of data that were used for the regression.

Contents

1 Background **1**
- 1.1 An outlook of the airline industry . . . 1
- 1.2 Purpose . . . 1
- 1.3 Questions targeted . . . 2
- 1.4 Scope . . . 2

2 Theory **3**
- 2.1 OLS - Ordinary Least Squares . . . 3
- 2.2 Assumptions . . . 4
- 2.3 Violations of the assumptions . . . 4
  - 2.3.1 Endogeneity . . . 4
  - 2.3.2 Heteroskedasticity . . . 5
  - 2.3.3 Multicollinearity . . . 6
  - 2.3.4 Additional violations . . . 7
- 2.4 Model selection . . . 7
  - 2.4.1 R^2, adjusted R^2 and η^2 . . . 8
  - 2.4.2 Akaike (AIC) . . . 8
  - 2.4.3 P-value and Confidence Intervals . . . 8
- 2.5 Variables . . . 9

3 Method **10**
- 3.1 Data manipulation . . . 10
- 3.2 Variables . . . 11
- 3.3 Model creation and validation . . . 11
  - 3.3.1 Response variable . . . 12
  - 3.3.2 Variable selection . . . 14
  - 3.3.3 Endogeneity . . . 14
  - 3.3.4 Heteroskedasticity . . . 15
  - 3.3.5 Multicollinearity . . . 15
  - 3.3.6 Linearity . . . 16
  - 3.3.7 General models . . . 20

4 Results **21**
- 4.1 Final Models . . . 21
- 4.2 Additional models . . . 27

5 Discussion **28**
- 5.1 Covariates . . . 28
- 5.2 Model . . . 30

6 Conclusion **32**

7 Industrial Engineering Approach **33**
- 7.1 Literature overview . . . 33
- 7.2 Customers . . . 34
  - 7.2.1 Consumer insights . . . 34
  - 7.2.2 Consumer behaviour . . . 35
- 7.3 The marketplace . . . 36
  - 7.3.1 The business model of an online travel agency . . . 36

7.3.2 The four P's of marketing in an online environment . . . 36
7.4 Communication . . . 38
7.4.1 Creating trust . . . 38
7.4.2 Channels . . . 38
7.5 Conclusion . . . 39

8 References **40**

9 Appendix **42**
9.1 Long distant destinations . . . 42
9.2 European destinations . . . 47

List of Figures

3.1 Histogram *before* log transformation BKK 12
3.2 Normal Q-Q plot *before* log transformation BKK 12
3.3 Histogram *after* log transformation BKK 13
3.4 Normal Q-Q plot *after* log transformation BKK 13
3.5 Histogram *before* log transformation LON 13
3.6 Normal Q-Q plot *before* log transformation LON 13
3.7 Histogram *after* log transformation LON 14
3.8 Normal Q-Q plot *after* log transformation LON 14
3.9 Scatter plot of *log*(*pricepers*) to *daysbetween* for BKK 17
3.10 Function of *daysbetween* compared to the observations 18
3.11 Scatterplot LON 19
3.12 Function of *daysbetween* compared to the observations 20
9.1 Scatter plot NYC 42
9.2 Scatter plot LAX 42
9.3 Scatter plot PAR 47
9.4 Scatter plot ROM 47

List of Tables

3.1 Multicollinearity test BKK 16
3.2 Multicollinearity test LON 16
3.3 Transformations of *daysbetween* examined 17
3.4 Transformations of *daysbetween* examined 19
4.1 Final result BKK 22
4.2 Weekday of the outbound flight, Sunday as benchmark 23
4.3 Weekday of the inbound flight, Sunday as benchmark 23
4.4 Departure airport, ARN as benchmark 23
4.5 Month of departure, January as benchmark 24
4.6 The number of days between booking and the outbound flight . . 24
4.7 Final result LON 25
4.8 Weekday of the outbound flight, Sunday as benchmark 26
4.9 Weekday of the inbound flight, Sunday as benchmark 26
4.10 Departure airport, ARN as benchmark 26
4.11 Month of departure, January as benchmark 26
4.12 The number of days between booking and the outbound flight . . 27
9.1 Final result NYC 43
9.2 Final result LAX 44
9.3 Weekday of the outbound flight for NYC and LAX, Sunday as benchmark 45
9.4 Weekday of the inbound flight for NYC and LAX, Sunday as benchmark 45
9.5 Departure airport for NYC and LAX, ARN as benchmark 45
9.6 Month of departure for NYC and LAX, January as benchmark . 45
9.7 The number of days between booking and the outbound flight for NYC and LAX 46
9.8 Final result PAR 48
9.9 Final result ROM 49

9.10 Weekday of the outbound flight for PAR and ROM, Sunday as benchmark 50
9.11 Weekday of the inbound flight for PAR and ROM, Sunday as benchmark 50
9.12 Departure airport for PAR and ROM, ARN as benchmark 50
9.13 Month of departure for PAR and ROM, January as benchmark . 50
9.14 The number of days between booking and the outbound flight for PAR and ROM 51

1 Background

1.1 An outlook of the airline industry

Looking at the airline industry today, it is rather different than it was a couple of years ago. In 1992, the Swedish air market was deregulated, which meant an increase in the destination selection, but also a significantly higher competition between airline companies and lower flight prices (Transportstyrelsen 2013). Due to this, a need for help finding the best airline tickets has arisen and a new market for so called online travel agencies, OTAs, appeared.

The emergence of the OTAs has generated new possibilities for consumers, in terms of comparing both different airline companies, destinations and time of the year for the trip. Customers may easily use the search functions on these OTAs, to find the best price for the journey of their choice, and then book the tickets through that site. There is also a business where search engines provide the similar search features and then redirect the customers to an OTA for the actual booking. However, things may not always be as straightforward. For instance, one might wish to find out how long time in advance it is optimal to book a trip for a specific destination, or at what time of the year it would be optimal to book the flight, in terms of minimizing the price. These kinds of problems arise as most online travel agencies and search engines only present the best price, given some specifications of the trip, that is available at that exact moment.

In a survey by the European Commission, it is stated that the total number of outbound trips in 2014 from Sweden by people at the age of 15 or older is a little above nine million (Eurostat 2015). There is a high competition between the OTAs and search engines for airline tickets and one way to compete is by providing the customers full service (Roger-Monzó, Martí-Sánchez, Guijarro-García 2015). However, there are different ways for an OTA to provide these services for its customers. One approach is by helping the customers to find the cheapest price for specific trips, where the potential findings of this study could be useful. Some search engines present similar services, such as Momondo and Skyscanner, in terms of guides on the different factors that have an impact on the price of the airline tickets. A regression analysis of the phenomenon yields an accurate result, which could work as the underlying for a booking guide to the customers. To further expand the employment of the mathematical analysis in this study, the marketing model of an OTA will also be examined, as to resolve how this study could be implemented in a marketing purpose.

1.2 Purpose

The underlying purpose of this study is to contribute to the growth of online travel agencies, by helping them to compete in the environment they operate in. This will be done by creating a tool that could be used in marketing purposes, in terms of providing customers with a booking guide, helping them to find the best price for a given trip. To achieve this, a regression analysis will be performed on the factors that might affect the prices of airline tickets. The purpose is not to produce a model for accurate price prediction, as the prices

of airline tickets are largely affected by private information that is difficult to predict, such as pricing strategies, that are not included in the model. Instead, the idea is to determine which factors have an impact on the prices of airline tickets, and how, to be able to answer the questions targeted.

As mentioned earlier, studies within the area have previously been conducted, although it is uncertain which methods were used to derive the results in those studies. No regression analysis performed on the area was found when performing the initial research for this study. This study will thus also examine whether it is even possible to perform a regression analysis on the area.

1.3 Questions targeted

In order to realize the purpose, the questions below will be targeted, for specific destinations:

- Which factors affect the price of airline tickets, and how?
- How many days in advance is it optimal to book a flight as to minimize the price?
- Does the weekday of the outbound flight, and the day of the the actual booking affect the price? If so, which day of the week is the cheapest?
- Are there any seasonal differences in terms of the price, and if so what time during the year is the cheapest to travel?
- Does the time of the day for the booking affect the price, and if so what time is the best to book?

1.4 Scope

This study is limited not to include factors such as oil prices. This is because of the fact that many airline companies hedge their exposure to the price fluctuations, making them less sensitive for the cost of jet fuel. Due to the high competition in the industry, airline companies find it hard to transfer their costs to their customers (Turner, Hoon Lim 2015). Combining these arguments, it was assumed that the oil prices would not be of any large relevance to the price of airline tickets. It was also determined that recent geopolitical events, nature disasters or terrorist attacks and alike would not be examined. The decision was based on conversations with Mattias Nyman, the CEO of Flygresor.se, who could provide his experience and knowledge of the industry and consumer behaviour.

2 Theory

Regression analysis is a method for describing the relation between a dependent variable, the response variable, and the independent variables that affect the outcome of the response variable, that is the covariates. There are different approaches to a regression analysis, where the model could be either structural or predictional and the data used when assessing the covariates can be either observational or experimental (Lang 2015, p. 1).

This study has been performed from a structural perspective. This perspective assumes that the covariates affect the value of the response variable and not the other way around, which does not have to be the case when performing a predictive interpretation, and then examines and tries to map this relation. The data on which the study was performed is observational, meaning that it was collected from events that we could not affect or interfere with.

2.1 OLS - Ordinary Least Squares

A linear regression model could be presented as below:

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_i \qquad i = 1, .., n \tag{2.1}$$

Where n is the number of observations, y_i the response variable, x_{ij} the value of covariate j and e_i the error term. The parameter β_j is unknown and is to be estimated from data.

The model shown in (2.1) can also be expressed in matrix notation, with Y and e as nx1 matrices and X as an $n\times(k+1)$ matrix, as:

$$Y = X\beta + e \tag{2.2}$$

The error term, or the residual, is a stochastic variable showing the difference between the actual value of the response variable and the estimation of it. The covariates are seen as deterministic (Lang 2015, p. 3). The model is said to be multiple when it includes more than one covariate, which is the case in this study. It is important to note that the model can only be used when five basic assumptions hold. These assumption will be presented in detail in section 2.2.

When the assumptions are met, the OLS estimator, Ordinary Least Squares, provides the best estimate for β (Kennedy 2008, p. 43). These estimates are denoted as $\hat{\beta}$ and represent the relation between the response variable and the covariates, where a value far from zero would indicate that the corresponding covariate has a strong influence on the value of the response variable.

The OLS estimate minimizes the sum of squared residuals, described mathematically as

$$\hat{e}^t\hat{e} = |\hat{e}|^2 \tag{2.3}$$

where

$$\hat{e} = Y - X\hat{\beta} \tag{2.4}$$

This is achieved by solving the normal equations (Kennedy 2008, p. 43).

$$X^t\hat{e} = 0 \tag{2.5}$$

From (2.4) and (2.5) the estimated β is

$$\hat{\beta} = (X^tX)^{-1}X^tY \tag{2.6}$$

2.2 Assumptions

First assumption

The dependent variable can be modelled as a linear function of the independent variables, plus an error term consisting of the residual (Kennedy 2008, p. 42). See equation (2.1) above.

Second assumption

The expected value of the error term is zero (Kennedy 2008, p. 42).

$$E(e_i) = 0 \tag{2.7}$$

Third assumption

All the error terms have the same variance and there exists no correlation between them (Kennedy 2008, p. 42).

$$E(e_i^2) = \sigma^2 \tag{2.8}$$

Fourth assumption

The covariates are considered to be deterministic, that is fixed in repeated samples (Kennedy 2008, p. 42).

Fifth assumption

No covariates are perfectly correlated with a linear combination of the others and there are more observations than number of covariates (Kennedy 2008, p. 42).

2.3 Violations of the assumptions

2.3.1 Endogeneity

Endogeneity among the observations will violate the second assumption. This means that one or several of the covariates are correlated with the residual, such that the expected value of the residual is no longer equal to zero (Lang 2015,

p. 25 - 26). When this occurs, the OLS-estimates will be inconsistent, such that they do not converge in probability to their true values. It is important to note that this phenomenon will only cause a problem if the interpretation of the regression model is structural (Lang 2015, p. 25 - 26).

Endogeneity can arise due to different circumstances, such as (i) a bias in the data selection, (ii) the dependent variable affecting the independent variables, as opposed to only having the opposite liaison (simultaneity), or (iii) incorrectly excluded covariates (Lang 2015, p. 25 - 28).

Detecting endogeneity

There is no prominent way of detecting endogeneity among the observations. The procedures taken in this study will be described more in detail in section 3.3.3 below.

Common measures

The measures vary depending on the origin of the endogeneity, but a common remedy is the usage of so called instrumental variables. This procedure aims to find exogenous variables, the instruments, that are highly correlated to the endogenous ones, though uncorrelated to the residual. The endogenous variables in the model are then replaced by the instruments.

2.3.2 Heteroskedasticity

The third assumption, as stated above, demonstrates homoskedasticity where the variance is the same for all the error terms. A heteroskedastic model results in the violation of this assumption, such that the variance of the error terms are not all the same. Through the past, it has been common to assume models as being homoskedastic, as this simplifies calculations. As stated earlier, the OLS-estimator provides the best estimates for a homoskedastic model, though it is inefficient for a heteroskedastic model (Lang 2015, p. 3, 7). If homoskedasticity is assumed when the model is in fact heteroskedastic, the point estimates will be the same, however their standard deviations will be greater (Lang 2015, p. 17).

Detecting heteroskedasticity

Two well known procedures used to detect heteroskedasticity are the Eye-ball Test and White's Variance Estimate (Kennedy 2008, p. 34). Nowadays, when essentially all computations are performed with computer software, assuming a heteroskedastic model will not cause any substantial computational difficulties. It is when falsely assuming a homoskedastic model that problems may arise, since this assumption includes the usage of estimators that are inefficient for heteroskedastic models. However, assuming a heteroskedastic model, the estimators used will provide accurate estimates for both types of models. Consequently, it is not necessary to verify whether the model is heteroskedastic, since it is possible to use estimators suited for both (Hansen 2016, p. 221), (Lang 2015, p. 3 - 4).

Common measures

A first remedy for heteroskedasticity is trying to make the error terms closer to homoskedastic, which can be done by reformulating the model or excluding outliers in the data. If heteroskedasticity would still be present, the estimator to use is White's Robust Estimate, where the estimated covariance matrix is given by:

$$C\hat{o}v(\hat{\beta}) = \frac{n}{n-k-1}(X^tX)^{-1}X^tD(e_i^2)X(X^tX)^{-1} \tag{2.9}$$

Here, as in previous formulas, n is the number of observations and k the number of covariates. The factor $\frac{n}{n-k-1}$ is an ad-hoc compensation in order not to underestimate the value of the residuals squared, and was not a part of the original estimate presented by White. Many computer programs used for regression analysis have a built in function for estimates for a heteroskedastic model, referred to as robust estimates (Lang 2015, p. 18).

2.3.3 Multicollinearity

When there is multicollinearity in the covariates, the fifth assumption is violated. This is since at least one of the covariates is dependent on a linear combination of some of the other covariates. The result is that these point estimates will have large variances, which causes difficulties when interpreting the model (Kennedy 2008, p. 193). Multicollinear covariates only affect the corresponding point estimates, not the estimates of the remaining covariates in the model. Because of this, there exists more than one school of thought, in terms of the seriousness of this phenomenon and when to take measures to adjust for it. Lang means that if the covariates explicitly examined are not multicollinear, it is irrelevant whether there are other covariates in the model that are multicollinear.

Detecting multicollinearity

There are several ways to detect multicollinearity, one of which is the Variance Inflation Factor (VIF), defined as:

$$VIF = \frac{1}{1-R^2} \tag{2.10}$$

R^2 is the coefficient of correlation, as explained in further detail in section 2.4.1, when running a regression on one specific covariate as dependent variable. Generally, if VIF > 10, there may be multicollinear covariates in the model (Lang 2015, p. 54 - 55). However, this model assumes that the covariates examined are continuous, and consequently it does not work on categorical variables, explained in section 2.5. There is, however, an application of the model that can be used on categorical variables, the Generalized Variance Inflation Factor (GVIF). GVIF can be used when the degrees of freedom of a covariate is more than one, as is the case for categorical covariates (Fox, Monette 1992). For continuous variables, GVIF and VIF are equivalent. For the categorical variables,

the measure $\text{GVIF}^{1/2Df}$ is calculated, showing the proportional decrease in precision of the estimated value due to collinearity. By squaring this measure, the VIF rule of thumb value could be used for collinearity on these variables as well.

Common measures

Alike endogeneity and heteroskedasticity, there are different procedures as to eliminate the errors caused by multicollinearity, where none is definite. For the categorical variables, the solution for avoiding multicollinearity within categories of that specific variable, is to remove one of the categories in the regression. Another way to eliminate multicollinearity is to increase the number of observations. It is also possible to remove one of the covariates, to eliminate multicollinearity. However, this may cause a bias in the model (Lang 2015, p. 15).

2.3.4 Additional violations

Another violation, that affects the first assumption, is when the relationship between the dependent variable and the covariates is not linear. The easiest way to detect this is to plot one variable at a time against the dependent variable, to see if a linear relationship can be distinguished. This is only done with the continuous variables. A nonlinear relationship can for instance be that one of the independent variables has a squared relation with the response variable. A measure in this case is then to add the concerned variable as a squared term to the model.

For the first assumption, a violation can also be the absence of relevant variables or the inclusion of unrelated variables. To escape this, it is appropriate to perform an Akaike test and calculate R^2, which will be described in section 2.4 below, including different covariates in the model (Lang 2015, p. 8, 21).

Further, the third assumption will be violated if the residuals are not uncorrelated, meaning that they are not normally distributed. This can be examined by performing residual plots.

2.4 Model selection

Generating a regression model that fits the problem targeted and also can be used on other sets of data, is not always evident. Finding this model, in terms of choosing which covariates to include or exclude, is the challenge in regression analysis. The inclusion or exclusion of the wrong covariates may lead to some of the assumptions stated above being violated, essentially leading to incorrect estimates. Here the inclusion of too many covariates may lead to the model being too well-fitted to the current dataselection. On the other hand, not including all essential covariates, may lead to a low level of explanation and a high residual. There are different approaches and measures to look at when finding a regression model, some of which will be described in this section.

2.4.1 R^2, adjusted R^2 and η^2

R^2 is a measurement of how good the model is, as it measures the extent to which the covariates explain the response variable (Lang 2015, p. 8). The R^2 of a model can be computed and interpreted in different ways, one of which is shown below.

$$R^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2} \tag{2.11}$$

Here, $\hat{e}$ is the sum of the residuals from the covariates in the model and $\hat{e}_*$ is the sum of the residuals when the regression is run on only the intercept, that is without the covariates. Consequently, the formula above shows the relative size of the part explained by the covariates. An important matter on model selection is, that by increasing the amount of covariates in the model, there is an associated cost in terms of the loss of degrees of freedom. To acknowledge this, adjusted R^2 can be used instead of R^2, as it accounts for the degrees of freedom lost when adding a covariate.

A similar employment of the method is to calculate the η^2 (eta squared), the effect-size. Eta squared works in an analogous way to the R^2, only it is a measure on the relative explanation of one or several of the covariates, as opposed to all of them (Lang 2015, p. 9).

2.4.2 Akaike (AIC)

As to choose which covariates to include in the model, there are several methods that can be employed, one being the Akaike Information Criterion test. The test is performed by computing the AIC-value for the full versus reduced model and then choosing the one that minimises this value.

$$AIC = n\ln(|\hat{e}|^2) + 2k \tag{2.12}$$

It is common to calculate the difference between the AIC-values of two models, containing different sets of covariates. If this difference is negative, then the reduced model is preferred (Lang 2015, p. 21-22). When evaluating covariates to include in the model, one should always keep in mind that Akaike does not provide a completely certain answer on which model is the best, and thus it should only be used as guidance. This is due to the fact that Akaike only assesses the issue of minimizing the information loss, which is why it is important to employ other measures as well when evaluating the model.

2.4.3 P-value and Confidence Intervals

The p-value is used for testing a hypothesis. In regression analysis, the hypothesis being tested, is that one or more of the beta values are equal to zero ($\beta_j^0 = 0$). The p-value is derived from the computations of the F-value, as presented in equation (2.13). The F-value is calculated under the null, which means that by assuming that the null hypothesis is true, the distribution of the F-value is

known.

$$F = \left(\frac{\hat{\beta}_j - \beta_j^0}{\mathrm{SE}(\hat{\beta}_j)} \right)^2 \tag{2.13}$$

The p-value is then the probability that X, which is a random variable with an F-distribution, is greater than the computed F-value.

$$p = Pr(X > F), \quad \text{where } X \in F(1, n-k-1) \tag{2.14}$$

First a level of significance is chosen, and if the computed p-value is less than this level, the hypothesis can be rejected, whereas if the p-value is higher than this, the hypothesis that the beta is equal to zero cannot be rejected at the chosen level of significance (Lang 2015, p. 8). The significance level is normally set to 5%.

A confidence interval, at level 1-$\propto$, for the beta values can be derived as shown below.

$$\hat{\beta}_j \pm \sqrt{F_{\propto}(1, n-k-1)}\mathrm{SE}(\hat{\beta}_j) \tag{2.15}$$

2.5 Variables

Variables where the data belongs to different categories, but the categories themselves do not have any specific values, are called categorical variables. In regression analysis, these types of covariates are managed by giving each observation the value 1 if the event that the covariate represents occurs, and 0 otherwise. When handling categorical variables, it is essential to remove one of the categories from the model, or else there would be multicollinearity. This is also important as to be able to compare the different categories to each other.

3 Method

The study is done by performing a regression analysis, using the programming language R. The layout of the study consists of (i) the collection and manipulation of data, (ii) the modeling of a well fitted and reliable model and (iii) the interpretation and analysis of the results. There are both pros and cons to this method, as to answer the questions targeted and realize the purpose. The most prominent advantage is the fact that the regression analysis is performed on real data, quantifying relations. Another evident advantage is that the regression model is designed to be transferable within the area, and could be used on other sets of data for specific destinations. This makes an efficient and quick tool that could be used for detecting future changes in the pricing of airline tickets. An important limitation is the fact that the reality is complex and very difficult to model mathematically. It is likely that problems dealing with multicollinearity, endogeneity and heteroskedasticity will arise.

This report focuses on six of the most popular destinations in the data attained from Flygresor.se. These six destinations are Bangkok, New York, Los Angeles, London, Paris and Rome. As it was assumed that the impact of the factors examined would vary between the destinations, one regression per destination was run. These destinations were specifically chosen among the most popular ones, as to be able to make the distinctions between the two categories of destinations in Europe and those more distantly located, thus three of each. Furthermore, it was examined whether it would be possible to attain a general model for these two categories, as they were assumed to show similar features in terms of the factors affecting the price. As Bangkok and London were the most popular destinations within each category, a model for each was generated and presented in the report. The models generated for the additional destinations were then compared to these two models, as to investigate the similarities and whether two general models could be attained. This will be further discussed in section 4.2.

3.1 Data manipulation

The regression was run on data from Flygresor.se, consisting of the lowest price available for searches on specific trips. The interval of the data ranged from May 2014 to May 2015, containing roughly 27.5 million observations. The range of the data collected was deemed to be appropriate, both as it includes seasonal changes, but also as the actual size of the initial data sample was large enough.

The primary data was saved in a .sql-file and due to the massive size of the file, a database in MySQL was created. The package RMySQL was used in order to communicate between the database and R, and to collect data from the database to use for the regression. First the data was filtered, and observations including specific features were excluded. Some of these features were to only include searches for adults, return trips and trips from the five largest airports of Sweden along with Copenhagen. Observations with searches done more than a year in advance than the outbound trip were also excluded at this stage. A random sample of 40% was then generated, approximately 7.2 million observations, mainly because it is challenging handling that large sets of data,

in an iterated process, such as a regression analysis.

3.2 Variables

The data provided was manipulated and transformed into the following variables:

Continuous variables

pricepers (response variable): the price of a roundtrip for one adult.

daysbetween: the number of days between the search date and the date of the outbound flight.

Categorical variables

searchinterval: the time of the day the search was conducted. This variable was divided into the four intervals, 00.00 - 06.00, 06.00 - 12.00, 12.00 - 18.00 and 18.00 - 24.00, where 00.00 - 06.00 was set as the benchmark.

depweekd: the weekday of the outbound flight where Sunday was set as the benchmark, similar for the covariates *bookweekd* and *homeweekd*.

bookweekd: the weekday of the search.

homeweekd: the weekday of the inbound flight.

from: the departure airport, where Stockholm Arlanda Airport (ARN) was set as the benchmark. The different airports are Bromma Stockholm Airport (BMA), Copenhagen Airport (CPH), Gothenburg Landvetter Airport (GOT) and Stockholm Skavsta Airport (NYO).

depmonth: the month of the outbound flight, where January was set as the benchmark.

holiday: if there was a holiday in Sweden during or starting two days after the outbound flight, where the event of no holiday was set as the benchmark. The holidays accounted for were winter break, Easter break, Ascension Day, Whitsuntide, autumn break and Christmas break.

3.3 Model creation and validation

Upon running the regressions for the different destinations, it was found that the data showed different relations between the covariates and the response variables, which is why particular models were generated for each of the destinations. The process of selecting a model is explained below, for the two destinations Bangkok (BKK) and London (LON), for which the results are then presented in section 5. This process was similar for the rest of the destinations, for which the results are presented in the Appendix.

3.3.1 Response variable

The natural logarithm of the variable *pricepers* was used as the response variable. The logarithm is used as to make the residuals closer to normally distributed, which is a common measure when the response variable is positive by nature and varies largely by size. It is important to note that when the natural logarithm of the response variable is used, the interpretation of the betas in the model changes from an absolute change in the value of the response variable, to a percentage change.

To check whether the residuals were normally distributed, a quantile-quantile-plot (qq-plot) was created, were the residuals are plotted against a hypothetical line, representing a normal distribution. If the residuals are normally distributed, they should follow the normal line, when plotted. To complement the interpretation of the qq-plot, a histogram was generated as well. The histogram illustrates the distribution of the residuals, which would be the classic bell-curve if they are normally distributed.

The plots and histograms for BKK and LON, presented below, show that the residuals converge significantly more towards a normal distribution after transforming *pricepers*.

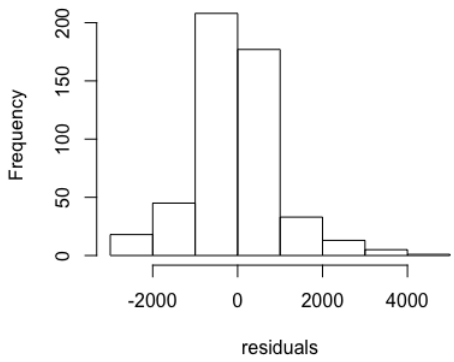

Figure 3.1: Histogram *before* log transformation BKK

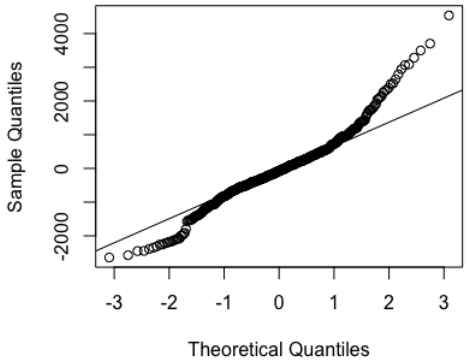

Figure 3.2: Normal Q-Q plot *before* log transformation BKK

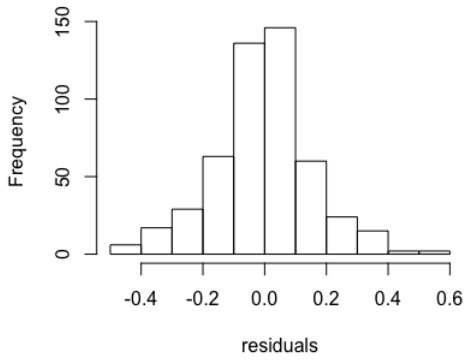

Figure 3.3: Histogram $after$ log transformation BKK

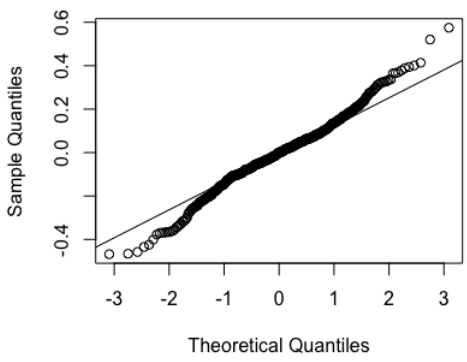

Figure 3.4: Normal Q-Q plot $after$ log transformation BKK

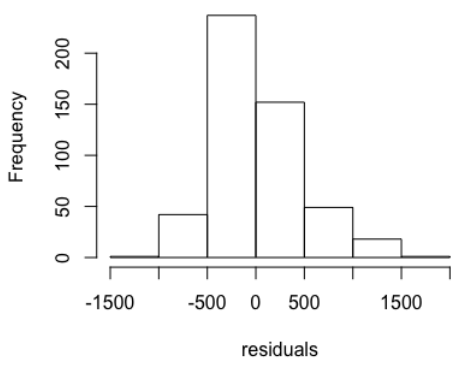

Figure 3.5: Histogram $before$ log transformation LON

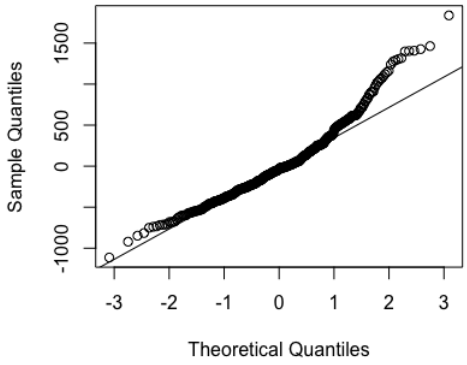

Figure 3.6: Normal Q-Q plot $before$ log transformation LON

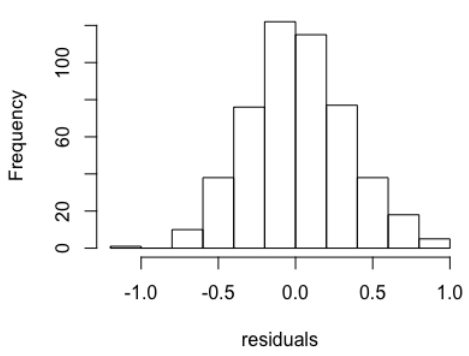

Figure 3.7: Histogram $after$ log transformation LON

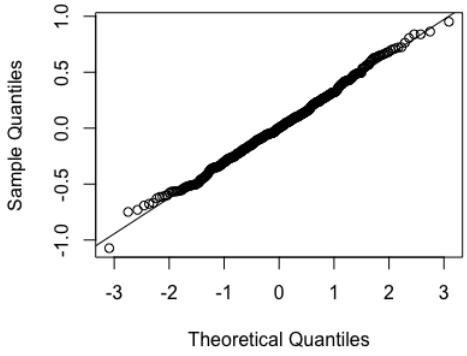

Figure 3.8: Normal Q-Q plot $after$ log transformation LON

3.3.2 Variable selection

Using the initial variables as a starting point, the covariates were examined in terms of their level of explanation (adjusted R^2), but also the p-value, Akaike and η^2, as to select which covariates to include in the final models. For each of the different measures, a level was set as a guidance when assessing the values, to determine whether to include that covariate or not. When actually assessing the values, they were put into context, and the measurements were weighted against each other. To include a covariate, the rules of thumb applied in this study were thus an increase in the R^2 to the full model, a p-value less than 5%, a positive Akaike-value and a η^2 greater than 1 permille.

BKK

After running a regression on the initial model, the covariates showing the most unfavorable values, *searchinterval* and *bookweekd*, were further evaluated. Although the two covariates both showed some slightly better and some slightly worse values for the three different measures, they were not convincing enough to be included in the model. In addition to this, both of the covariates' confidence intervals contained zero.

LON

A regression on the initial model for London generated similar results to those of Bangkok, where the covariates *searchinterval* and *bookweekd* showed weak values. Accordingly, the same arguments were applied for the London model and the two covariates were excluded.

3.3.3 Endogeneity

A formal test of examining whether there was endogeneity in the model was not conducted, see section 2.3.1. However, the issue has been addressed. As for

the data selection, the initial set covered all observations in a full year, from which a random selection was made. Due to these factors, there does not seem to be any bias in the data selection. One factor that potentially could have caused endogeneity is the fact that the data used in the regression is from a specific company, meaning that it only includes the airline companies that the company works with. However, as the company in question is rather large and works with many different airline companies, this was not considered to cause any endogeneity.

Regarding simultaneity, it was concluded that this did not exist in the model, seeing as a shift in the dependent variable would not necessarily imply a shift in any of the covariates. As to support this argument, a comparison with models demonstrating a supply and demand relation, which are commonly used as an example when simultaneity might occur, was made. In these examples, simultaneity occurs since the dependent variable demand has an impact on the covariate price. A similar relation was not found in our model, and thus the conclusion presented was drawn.

As for the exclusion of covariates, this study has a clear scope (section 1.4), so the factors excluded from the beginning have been thoroughly evaluated, and should therefore not affect the validity of the model. The selection of variables is also studied further, assuring that the covariates included in the model have a high level of explanation.

3.3.4 Heteroskedasticity

The presence of heteroskedasticity was not reviewed, instead the function $robust.summary()$ was used to calculate robust estimates.

3.3.5 Multicollinearity

The existence of multicollinearity was examined at an early stage. One of the questions addressed was how many days in advance it would be cheapest to book a flight. To attain this covariate, *daysbetween*, the difference between the search date and the date of the outbound flight was calculated. As *daybetween* is then a linear combination of the other two variables, it was decided not to include the covariate for the search date.

Looking at the date of the outbound flight, the covariate *depmonth* was divided per month. This was as to avoid the multicollinear relation that would have arisen between this covariate and the covariates *depweekd* and *holiday*, had the departure date been divided per day or week.

GVIF BKK

The function $vif()$ from the car-package was used to calculate the GVIF and $\text{GVIF}^{1/2Df}$ for the covariates. The values were well below the limit of 10, and thus there should be no multicollinearity in the model.

	GVIF	Df	GVIF^(1/(2*Df))
depweekd	1.07	6.00	1.01
homeweekd	1.02	6.00	1.00
from	1.01	4.00	1.00
depmonth	1.42	11.00	1.02
daysbetween	1.08	1.00	1.04
holiday	1.33	1.00	1.15

Table 3.1: Multicollinearity test BKK

GVIF LON

Performing the GVIF test on the London-model resulted in values considerably less than 10 as well, so the model does not seem to have any multicollinearity.

	GVIF	Df	GVIF^(1/(2*Df))
depweekd	1.20	6.00	1.02
homeweekd	1.17	6.00	1.01
from	1.04	4.00	1.00
depmonth	1.73	11.00	1.03
daysbetween	1.16	1.00	1.08
holiday	1.42	1.00	1.19

Table 3.2: Multicollinearity test LON

3.3.6 Linearity

The linearity of the model is examined by looking at the relation between the continuous covariates and the response variable, in this case only *daysbetween*. The two were plotted against each other and a visual interpretation can then be done, as to assess the relation between them.

BKK

The plot for BKK is shown below in figure 3.9.

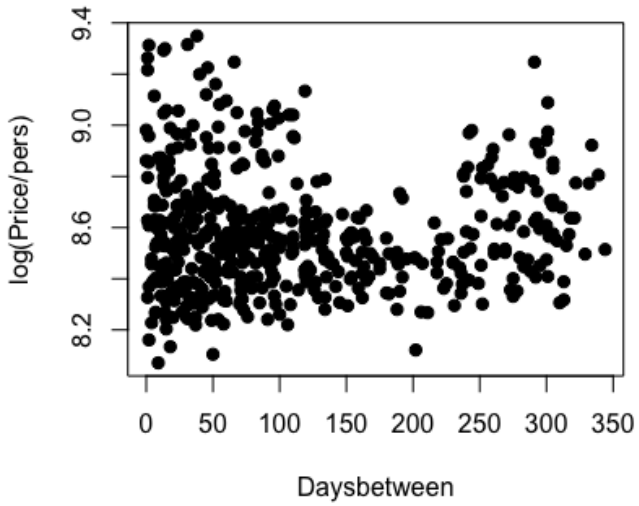

Figure 3.9: Scatter plot of $log(pricepers)$ to *daysbetween* for BKK

Showing the classic smiley, the relation seems to have square features, although somewhat skewed towards the left, showing similar features to those of a logarithmic curve. Consequently, the square, natural logarithm and natural logarithm squared of *daysbetween* were added to the model. To determine which of the variables would increase the linearity of the model the most, their Akaike-, p-, η^2- and R^2-values were examined. All the added covariates showed similar values, meeting all the pre-set levels for the different measures, which is why it was decided to include all of them in the model.

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
daysbetween	-0.0028	0.0000	0.0061	0.0000	-0.0029	-0.0027
squaredays	0.0000	0.0000	0.0065	0.0000	0.0000	0.0000
logdays	-0.1687	0.0040	0.0110	0.0000	-0.1766	-0.1608
I(logdays^2)	0.0324	0.0009	0.0064	0.0000	0.0306	0.0343

Table 3.3: Transformations of *daysbetween* examined

$$\begin{aligned} log(pricepers) =& 8.794199 - 0.002797x + 0.000005x^2 - 0.168704\ln(x) + \\ & 0.032440\ln^2(x) \end{aligned} \tag{3.1}$$

The function above consists of the response variable depending on the included transformed covariates. In order to validate the inclusion of the transformed

covariates, the graph of the function was plotted and compared to the scatter plot. The sharp decline in the graph when there is only a few days between the booking and the outbound flight incorporates the cluster of points in the scatter plot in the lower left. Likewise, the dip in the graph around 200 days is similar to the pattern of the points in the scatter plot.

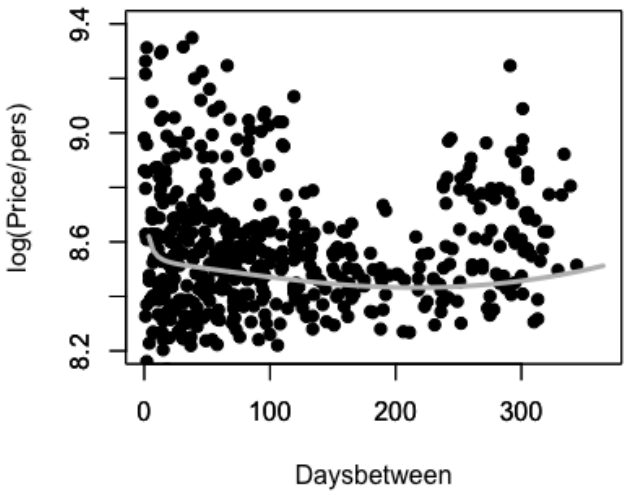

Figure 3.10: Function of *daysbetween* compared to the observations

LON

To evaluate the relation of the covariate *daysbetween* and *pricepers* in the London model, the same scatter plot was used as for Bangkok, presented below in figure 3.11.

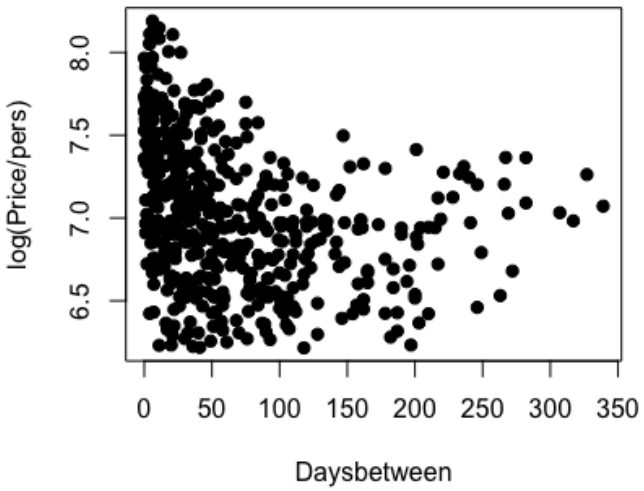

Figure 3.11: Scatterplot LON

The relation seems to be logarithmic, considering the band of points in the far left and then the dispersion moving in the right direction. The scatter plot could also be interpreted as if the graph would contain some square features. Thus, all three transformations of *daysbetween* were added to this model as well, for further analysis. Looking at the measures, the pre-set levels for all measures were met only for *daysbetween* and *logdays*, making it clear that they should be included in the final model. The p-value of *squaredays* may meet the level of less than 5%, but its η^2 is very low, and the R^2 of the model is not improved when adding the covariate. Hence, it was decided not to include *squaredays* or *logsquaredays* in the model.

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
daysbetween	0.0017	0.0002	0.0006	0.0000	0.0013	0.0021
squaredays	-0.0000	0.0000	0.0000	0.0361	-0.0000	-0.0000
logdays	-0.2079	0.0067	0.0107	0.0000	-0.2211	-0.1947
I(logdays^2)	-0.0023	0.0020	0.0000	0.2330	-0.0062	0.0015

Table 3.4: Transformations of *daysbetween* examined

$$log(pricepers) = 7.765830 + 0.001295x - 0.212773\ln(x) \tag{3.2}$$

The function above consists of the response variable depending on the included transformed covariates. The same approach as for the Bangkok model was used

in order to validate the inclusion and exclusion of covariates. The graph seems well fitted, as the decline in the beginning accounts for the bundle of points in the scatter plot, and then the graph follows the slight increase in price, as illustrated.

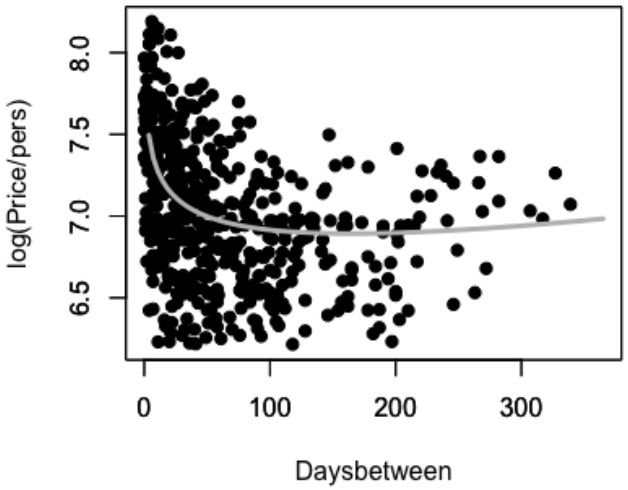

Figure 3.12: Function of *daysbetween* compared to the observations

3.3.7 General models

In order to examine whether a general model could be attained for each of the two categories, European cities and distant located cities, the models generated for the four additional destinations were compared to those of Bangkok and London. The intuition behind this, was that if the models within respective category would turn out to be similar, this could imply that the Bangkok and London models could be directly applied in a more general sense, that is for all destinations that would fall into either the two categories. The same procedure as described throughout this section was carried out when finding suitable models for the four complementary destinations.

4 Results

4.1 Final Models

BKK

The final model for Bangkok is presented below. A specification of the covariates is presented in table 4.1.

$$\begin{aligned} log(pricepers) =& \beta_0 + \beta_1 depweekd + \beta_2 homeweekd + \beta_3 from + \beta_4 depmonth + \\ & \beta_5 daysbetween + \beta_6 squaredays + \beta_7 logdays + \beta_8 logdays^2 + \\ & \beta_9 holiday \end{aligned} \tag{4.1}$$

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	8.79	0.00	0.95	0.00	8.78	8.80
depweekdMon	-0.02	0.00	0.00	0.00	-0.02	-0.02
depweekdTues	-0.02	0.00	0.00	0.00	-0.02	-0.02
depweekdWed	-0.01	0.00	0.00	0.00	-0.01	-0.01
depweekdThurs	0.02	0.00	0.00	0.00	0.02	0.02
depweekdFri	0.05	0.00	0.01	0.00	0.05	0.05
depweekdSat	0.03	0.00	0.00	0.00	0.02	0.03
homeweekdMon	-0.04	0.00	0.00	0.00	-0.04	-0.03
homeweekdTues	-0.03	0.00	0.00	0.00	-0.03	-0.03
homeweekdWed	-0.05	0.00	0.01	0.00	-0.05	-0.04
homeweekdThurs	-0.03	0.00	0.00	0.00	-0.03	-0.03
homeweekdFri	-0.01	0.00	0.00	0.00	-0.01	-0.01
homeweekdSat	0.03	0.00	0.00	0.00	0.03	0.03
fromBMA	0.14	0.00	0.01	0.00	0.13	0.14
fromCPH	0.07	0.00	0.03	0.00	0.07	0.07
fromGOT	0.12	0.00	0.08	0.00	0.12	0.12
fromNYO	0.75	0.12	0.00	0.00	0.51	0.99
depmonth(1,2]	-0.09	0.00	0.02	0.00	-0.09	-0.09
depmonth(2,3]	-0.06	0.00	0.01	0.00	-0.06	-0.06
depmonth(3,4]	-0.14	0.00	0.04	0.00	-0.14	-0.14
depmonth(4,5]	-0.21	0.00	0.04	0.00	-0.21	-0.21
depmonth(5,6]	-0.04	0.00	0.00	0.00	-0.05	-0.04
depmonth(6,7]	0.14	0.00	0.03	0.00	0.14	0.14
depmonth(7,8]	-0.05	0.00	0.00	0.00	-0.06	-0.05
depmonth(8,9]	-0.13	0.00	0.01	0.00	-0.13	-0.13
depmonth(9,10]	-0.12	0.00	0.03	0.00	-0.13	-0.12
depmonth(10,11]	-0.09	0.00	0.02	0.00	-0.10	-0.09
depmonth(11,12]	0.23	0.00	0.17	0.00	0.23	0.24
daysbetween	-0.00	0.00	0.01	0.00	-0.00	-0.00
squaredays	0.00	0.00	0.01	0.00	0.00	0.00
logdays	-0.17	0.00	0.01	0.00	-0.18	-0.16
I(logdays^2)	0.03	0.00	0.01	0.00	0.03	0.03
holiday	0.13	0.00	0.11	0.00	0.13	0.13

Table 4.1: Final result BKK

The adjusted R^2 for the final model is 0.5735, which means that 57.35% of the variation in the price is explained by the model. It was also shown that none of the confidence intervals contained zero. Consequently, the hypothesis that the covariates have no impact on the price is rejected, at a confidence level of 95%.

Residual standard error: 0.153
Degrees of freedom: 357938
Multiple R^2: 0.5735
Adjusted R^2: 0.5735

The tables below provide a breakdown of how much the price differs in percentage, from the respective benchmark. As the response variable is in logarithm-form, the percentage change in its value can be obtained by the formula

$$100(e^{\beta} - 1) \tag{4.2}$$

A negative sign indicates a lower price and no sign indicates a higher price.

Weekday	Difference (%)
Monday	-2
Tuesday	-2
Wednesday	-1
Thursday	2
Saturday	3
Friday	5

Table 4.2: Weekday of the outbound flight, Sunday as benchmark

Weekday	Difference (%)
Monday	-4
Wednesday	-4
Tuesday	-3
Thursday	-3
Friday	-1
Saturday	3

Table 4.3: Weekday of the inbound flight, Sunday as benchmark

Airport	Difference (%)
CPH	7
GOT	13
BMA	15
NYO	112

Table 4.4: Departure airport, ARN as benchmark

Month	Difference (%)
May	-19
April	-13
September	-12
October	-12
February	-9
November	-9
March	-6
August	-5
June	-4
July	15
December	26

Table 4.5: Month of departure, January as benchmark

As the covariate *daysbetween* is continuous, the interpretation is somewhat different than the covariates presented above. It is nonetheless possible to find the number of days between booking and the outbound flight that yields the lowest price, by finding the minimum point of the function. The results are presented below.

	Number of days
lowest price	216

Table 4.6: The number of days between booking and the outbound flight

LON

The final model for London is presented below.

$$\begin{aligned} log(pricepers) =& \beta_0 + \beta_1 depweekd + \beta_2 homeweekd + \beta_3 from + \\ & \beta_4 depmonth + \beta_5 daysbetween + \beta_6 logdays + \beta_7 holiday \end{aligned} \tag{4.3}$$

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	7.77	0.01	0.94	0.00	7.75	7.78
depweekdMon	-0.05	0.00	0.00	0.00	-0.05	-0.04
depweekdTues	-0.04	0.00	0.00	0.00	-0.05	-0.04
depweekdWed	-0.00	0.00	0.00	0.63	-0.01	0.01
depweekdThurs	0.11	0.00	0.01	0.00	0.10	0.12
depweekdFri	0.05	0.00	0.00	0.00	0.04	0.05
depweekdSat	0.04	0.00	0.00	0.00	0.04	0.05
homeweekdMon	-0.10	0.00	0.01	0.00	-0.11	-0.10
homeweekdTues	-0.19	0.00	0.03	0.00	-0.20	-0.18
homeweekdWed	-0.21	0.00	0.04	0.00	-0.21	-0.20
homeweekdThurs	-0.20	0.00	0.03	0.00	-0.20	-0.19
homeweekdFri	-0.15	0.00	0.02	0.00	-0.15	-0.14
homeweekdSat	-0.07	0.00	0.00	0.00	-0.07	-0.06
fromBMA	0.20	0.01	0.01	0.00	0.19	0.22
fromCPH	-0.34	0.00	0.14	0.00	-0.35	-0.34
fromGOT	-0.08	0.00	0.01	0.00	-0.09	-0.08
fromNYO	-0.37	0.00	0.19	0.00	-0.38	-0.37
depmonth(1,2]	-0.23	0.01	0.02	0.00	-0.24	-0.22
depmonth(2,3]	-0.06	0.01	0.00	0.00	-0.08	-0.05
depmonth(3,4]	0.07	0.00	0.00	0.00	0.06	0.08
depmonth(4,5]	-0.03	0.01	0.00	0.00	-0.04	-0.02
depmonth(5,6]	0.01	0.01	0.00	0.02	0.00	0.02
depmonth(6,7]	0.14	0.01	0.01	0.00	0.13	0.15
depmonth(7,8]	0.17	0.01	0.01	0.00	0.16	0.18
depmonth(8,9]	0.08	0.01	0.00	0.00	0.07	0.09
depmonth(9,10]	0.11	0.01	0.01	0.00	0.10	0.12
depmonth(10,11]	0.08	0.01	0.00	0.00	0.07	0.09
depmonth(11,12]	0.31	0.00	0.05	0.00	0.30	0.32
daysbetween	0.00	0.00	0.03	0.00	0.00	0.00
logdays	-0.21	0.00	0.21	0.00	-0.22	-0.21
holiday	0.33	0.00	0.18	0.00	0.32	0.33

Table 4.7: Final result LON

The adjusted R^2 for the London model is 58.91%. In this model, only one of the covariates' confidence interval contained zero, namely *depweekdWed*, and thus it is not possible to determine that this covariate would be of significance to the price. However, none of the other confidence intervals contained zero, and for these covariates the hypothesis that the covariates have no impact on the price is rejected, at a confidence level of 95%.

Residual standard error: 0.281
Degrees of freedom: 95598
Multiple R^2: 0.5892
Adjusted R^2: 0.5891

The tables below provide a breakdown of how much the price differs, in percentage, from the respective benchmark. A negative sign indicates a lower price and no sign indicates a higher price.

Weekday	Difference (%)
Monday	-4
Tuesday	-4
Wednesday	0
Friday	5
Saturday	5
Thursday	11

Table 4.8: Weekday of the outbound flight, Sunday as benchmark

Weekday	Difference (%)
Wednesday	-19
Thursday	-18
Tuesday	-17
Friday	-14
Monday	-10
Saturday	-6

Table 4.9: Weekday of the inbound flight, Sunday as benchmark

Airport	Difference (%)
NYO	-31
CPH	-29
GOT	-8
BMA	23

Table 4.10: Departure airport, ARN as benchmark

Month	Difference (%)
February	-21
March	-6
May	-3
June	1
April	7
September	8
November	9
October	12
July	15
August	18
December	37

Table 4.11: Month of departure, January as benchmark

The interpretation of *daysbetween* and its transformations is analogous to that in the Bangkok model, as was the method for finding the number of days between booking and the outbound flight, yielding the lowest price. The result is presented below.

	Number of days
lowest price	164

Table 4.12: The number of days between booking and the outbound flight

4.2 Additional models

For the distant located complimentary destinations, New York and Los Angeles, the scatter plots showed a similar relation between $log(pricepers)$ and the covariate *daysbetween* to that of Bangkok. Based on this, it was deemed likely that a model similar to that of Bangkok would suit these two destinations. Looking at the initial covariates, the same ones were included as in the Bangkok model. Seeing that the values for inclusion or exclusion of the transformations of *daysbetween* all aligned well with the pre-set levels, the same ones were included, as in the Bangkok model. It was therefore concluded that the Bangkok model was applicable to the two destinations examined. Based on this, it is probable that the Bangkok model can be applied to other major distant located destinations. It is important to note that it is only the model that is the same for the destinations within the category, the estimates for the beta-values are specific for each destination. These are presented in the Appendix.

The same intuition was applied when examining the two European destinations, Paris and Rome. As their scatter plots showed similar relations between $log(pricepers)$ and the covariate *daysbetween* to that of London, the same procedure was taken, and it was found that the inclusion of the initial covariates was the same as for the London model. Looking at these values for the transformations of *daysbetween* however, the inclusion or exclusion of covariates was not completely aligned to the pre-set levels. These discrepancies were only marginal however, which is why it was concluded that the model created for London was applicable for the two complimentary destinations as well. Seeing as there were only minor differences in the values of the measures amongst the three destinations, it is probable that the London model could be applicable to other major European cities.

5 Discussion

In the discussion of the covariates, it is assumed that the price is relative to the demand, where a high demand would entail higher prices for the flights during that period.

5.1 Covariates

depweekd

The most expensive weekdays, Thursday through Saturday, are the same for the two destinations Bangkok and London. These results are considered as reasonable, seeing as people usually leave for a vacation in the second part of the week. There is however, a significant difference between the most expensive days, where outbound flights to London on Thursdays are much more expensive, than the rest of the days. As in line with the previous argument, this is also reasonable, as trips to London are usually of the shorter type long weekends, starting later in the week, mainly Thursdays. This coordinates with the fact that the first days of the week are cheaper for the outbound flight. As Bangkok is further away than London, trips there usually last longer, making the day of the outbound flight slightly less important.

homeweekd

For the inbound flight from Bangkok, Saturdays and Sundays are the most expensive, and Saturday through Monday for London. The results are deemed to be reasonable for both destinations, but due to somewhat different reasons. As for Bangkok, it is likely that the dates are coordinated with the hotel preference, where hotels may have special deals for accommodation on a weekly basis, which often begin and end during weekends. For trips to London, it was previously assumed that a great part of those would be long weekends, where it is common to fly back either in the end of the week, especially Sundays, which are significantly more expensive.

from

For this variable, the most expensive versus the cheapest departure airports differed significantly between the two destinations. Leaving for Bangkok, Arlanda is the cheapest airport and Skavsta is the most expensive. After discussing this with Mattias Nyman, the results seem accurate. For destinations at a longer distance, such as Bangkok, there are more flights leaving from Arlanda, whereas flights from Skavsta would include a change of flights. When leaving for London, the situation is almost reverse, where Skavsta is the cheapest airport and Bromma is the most expensive. These results were also discussed with Mattias Nyman, and it was concluded that as there are many low-fare airlines flying to European destinations from Skavsta, this relation was also accurate.

depmonth

It is more expensive travelling to Bangkok in January, July and December, than

the rest of the year. This aligns with the peak seasons in Thailand, which are located in connection to holidays, lasting longer than a couple of days or a week. When travelling to London, February is much cheaper than the rest of the year, whereas December is much more expensive than the rest of the year. This coordinates with the seasonal variations one might expect for a major European city.

holiday

For both destinations, it was more expensive to travel during a holiday, especially for London. This might be due to the fact that trips to closely located destinations would often occur on the shorter holidays, whereas trips to more far located destinations would be more dependent on seasonal variations.

daysbetween

As the covariate *daysbetween* is continous, it allows for further investigation, than the categorical covariates. As the number of days between booking and the outbound flight was attained from the graph created, there is some uncertainty in the result. This is because of the difficulty in fitting a curve to a scatter plot. In this case, the spreads in the prices on a single day were of substantial size, making it difficult finding a graph that would account for all these points. However, the graph provides an estimate of how many days in advance the lowest price is attained.

Studying the graph of the function in the Bangkok model, it is clear that the price drastically increases only when it is one or two days left between the booking and the outbound flight. It is also noted that the flat decrease in price as the number of days increases, before the minimum price at 216 days, is approximately the same as the increase in the price as the number of days increases, after the minimum. However, due to the flatness of the curve, meaning that the price does not change by much around this interval of days, there is additional uncertainty to the results, as it is difficult finding an exact number of days. The interpretation should thus be that the lowest price could be attained by booking *around* 216 days before the outbound flight.

In the London model, the drastic increase occurs when booking around ten days or less prior to the outbound flight. The same arguments regarding uncertainty in the results as in the Bangkok model should be applied here, as the band of points, that is a large spread in prices, is difficult to account for when fitting a graph. After that interval the curve is rather flat, meaning that the price does not change by much, apart from the minimum price at 164 days.

It important that the covariate *daysbetween* should be given a general application, considering its complex interpretation. For all the three distant located destinations, the graphs were rather similar, and given the somewhat square relation between the covariate and the price, it is possible to determine a minimum point in the curves, that is the number of days yielding the lowest price. Looking at the three European destinations however, the relations tended towards a logarithmic curve, making their graphs more difficult to interpret, in terms of finding a specific number of days. As stated above, the number of

days for the lowest price was computed for the London model, but the graphs for the Paris and Rome models did not show the slight increase towards the far right, as the London model did. An actual number of days was attained for Paris, but not Rome. However, there is uncertainty in the results, as both graphs indicate that the prices decrease as the number of days increase. Thus, the main takeaway should not be the specific number of days between booking and the outbound flight that yields the lowest price, but rather the increase in prices as the number of days approaches zero.

searchinterval and bookweekd

In both the models, the covariates *searchinterval* and *bookweekd* were excluded. They were initially thought to be of a rather high significance, hence the inclusion in the questions targeted. However, the weak values rejected this view. Removing the covariate *searchinterval* also makes sense from a contextual view. It is probable that the intraday price changes are only marginal, especially compared to the other covariates, showing price changes over longer periods. Similar arguments were applied to the removal of *bookweekd*, as it is reasonable that the price does not vary by much depending on the weekday of the booking, but rather over months and other seasonal differences.

5.2 Model

In this study, the data used for the regression only ranged over a year. As stated above, it was considered to be sufficient for the scope of the study. Nonetheless, regressing on data that ranges over several years would incorporate differences between years. An example of this would be if a cold summer one year would increase the demand for trips the subsequent summer, which could thus imply an increase in the price for those trips.

The R^2 of the two models are both around 60%, which was considered as quite high. Still, it is possible that these values might have been higher, had alternative covariates been included in the model. However, as stated in the Purpose in section 1.2, the idea of this study was not to compute a model for prediction of the prices of airline tickets, it was mainly to investigate some interesting factors, and their impact on the price. Given this approach, the exclusion of covariates, resulting in the lower R^2-values, would not impact the results as significantly, as if the purpose had been to compute a model for prediction of the prices. In this study, it is however still possible that the potential lack of relevant covariates has affected the estimates of the beta-values, making them slightly less accurate.

Regarding the general applications of the models generated for Bangkok and London, the results attained were not indisputable. The intuition was that destinations could be bundled together, based on the distance. As explained in the results section, the Bangkok and London models could be applied to the additional respective destinations, and thus the conclusion was drawn that the two models could probably work as proxies for similar destinations. However, as each model was only compared to two additional ones, the conclusion is not completely reliable. In order to attain the best possible model, a regression on

each destination should be run. However, given the scope of this study and report, that was not feasible. It is suggested that this should be done in future research, in order to be able to conduct a fair comparison of the models, to conclude whether the same model could be used for a whole category of destinations. Accordingly, the idea of mainly presenting the models for the two destinations Bangkok and London in this report was to attain two well fitted models, from which the principles and method could be applied to additional destinations.

6 Conclusion

Going back to the questions targeted in this study, the results occasionally differed among the destinations. Therefore, a general answer is provided when possible and otherwise, the results for Bangkok and London will be presented.

- What factors affect the price of airline tickets, and how?

For the destinations examined, the factors found to be of relevance to the price are the weekday of the outbound and inbound flight, the departure airport, the month of the outbound flight, the number of days between the booking and the outbound flight, and finally, if there was a holiday during the trip. The impact of the factors differed among the destinations and it is thus not possible to give a general answer to this question.

- How many days in advance is it optimal to book a flight as to minimize the price?

For flights to Bangkok, the lowest price will be attained by booking the trip around 216 days prior to the outbound flight, and respectively, around 164 days for London.

- Does the weekday of the outbound flight, and the day of the the actual booking affect the price? If so, which day of the week is the cheapest?

It was found that the day of the booking did not have a significant impact on the price, for either destination. However, as for the day of the outbound flight, is was found that the first half of the week was the cheapest for both the destinations, whereas Friday was the most expensive day to leave for Bangkok and Thursdays for London. The internal differences between the days were greater for London than for Bangkok.

- Are there any seasonal differences in terms of the price, and if so what time during the year is the cheapest to travel?

For respective destination, there are differences in the price between different months of departure. For Bangkok it was found that May was the cheapest month, and December the most expensive. Travelling to London is the cheapest in February, and most expensive in December.

- Does the time of the day for the booking affect the price, and if so, what time is the best to book?

The time of the day for the booking does not have an impact on the price, for either destination.

To conclude the findings, it is possible to minimize the price for a trip with some specific attributes, for all the six destinations examined, where similarities were found within the two respective categories.

7 Industrial Engineering Approach

Selling a product or service does not only evolve around the product. In today's modern society, there exists a substitute for most products and services available, and there is a high competition between retailers. To be able to communicate with the customers, raising awareness about the gods that are to be sold is essential. In order to do this however, it is imperative to understand the customers' behaviour, the best way of communicating to them and how different marketplaces work, only to mention a few. But as society has changed, so have the rules of marketing. Specifically, online retailers may face different challenges, and rewards, compared to the classic physical stores.

This section of the report will focus on how a marketing model could look, for online retailers, specifically online travel agencies, also known as OTAs. The approach will be a qualitative study, focusing on principles and theories from the textbook Principles of Marketing, by Philip Kotler and Gary Armstrong. The focus of the study will be (i) customers, (ii), the marketplace and (iii) communication. Analyzing OTAs and their specific marketing needs from these perspectives, exhibits how classical marketing models and procedures could be applied for an OTA, and demonstrates how the concept of regression analysis could be used in marketing purposes.

7.1 Literature overview

According to Kotler and Armstrong, one of the main purposes of marketing is to build solid relationships with the customers. In the textbook Marketing Management Strategies, by O.C. Ferrell and Michael D. Hartline, the authors build a great part of their analysis on this aspect. A cornerstone for this approach is the fact that, according to themselves, marketing has shifted towards an environment where the customers are in control. As online services, both social and functional, have become a greater part of the lives of many people, consumers are able to research competitors offering the same type of products or services (Ferrell, Hartline 2011, p. 5). Although this has put a pressure on retailers, especially online retailers, it has also enabled for a new type of marketing, where a two-way relation with the customers is the focus (Kotler, Armstrong 2012, p. 17). In his book The New Rules of Viral Marketing, the marketing strategist David Meerman Scott introduces the concept of *word-of-mouse*. It is an interpretation of the classic concept of word-of-mouth in an online environment. He suggests that online marketing has an organic approach, where companies and commercials are spread through recommendations, as opposed to impersonal marketing being exposed to whole segments of customers. Ferrell and Hartline share this view, as they argue that technology has even helped companies target specific customer groups, and create personalized marketing (Ferrell, Hartline 2011, p. 20). However, the authors claim that privacy and confidentiality problems arise when retailers are given the possibilities in using technology to target customers. Going back to the relation part of marketing, it is thus imperative that retailers act in a responsible way when interacting in an online environment with their customers.

7.2 Customers

7.2.1 Consumer insights

By gathering data from different parts of the business and sort it based on different categories, retailers can easily access this internal data and structure their marketing accordingly (Kotler, Armstrong 2012, p. 101). This is particularly convenient for online retailers, as all information is immediately digital, and thus accessible. A great benefit of these types of databases, as the authors argue, is the possibility of targeting consumers, and personalize the marketing.

To evolve this method further, retailers also conduct market researches. In the book, the two forms of data, secondary data, which is data that already exists, and primary data are presented. The primary data is gathered by conducting some sort of research, such as observational, experimental research or through surveys (Kotler, Armstrong 2012, p. 105 - 110). As described, observational data is collected by observing patterns, and customers' behaviour. The book prompts surveys as a diverse way of gathering information, given the level of accuracy that can be achieved, by adjusting both the questions, but also the way the survey is conducted, in terms of email, telephone or personal interviews (Kotler, Armstrong 2012, p. 110). For an OTA, the results from the regression analysis run in this study, is an example of how to use secondary data for marketing purposes. Although the regression is not directly run on consumer preferences, it is a linkage to consumer behaviour, in terms of when customers search for and book flights to different destinations. The results from the regression can be used as a guide to the customers, providing them with insight, based on real data, on how to find the lowest prices of airline tickets for specific destinations. Not only would the accuracy of this guide be an appealing factor to the customers, but it might also be a way of creating trust, by ensuring customers that their information stored, can be used to directly favour themselves.

A most important aspect that the book emphasizes is the quandary of consumers unwilling to share information, due to privacy reasons (Kotler, Armstrong 2012, p. 124). As stated above, the collection and analysing of consumer data enriches the possibilities of personalized marketing. In the highly competitive business where the OTAs operate, a "one size fits all"-concept might not be efficient enough as to capture the attention of the customers. The McKinsey report *iCOnsumers: Life Online* states that 35% of the people targeted are willing to share personal information in order to receive personalized marketing (McKinsey, p. 12). But what about the other 65%? There is a thin line between personalising the content of marketing, and breaching integrity. Although customers enjoy personalized marketing, suited for them, retailers do not all play by the same rules. For brands with a wide recognition, and well known product and services, customers tend to appreciate personalized marketing, but as for less known brands, consumers are less keen on receiving the same type of personalized marketing (Bleier, Eisenbeiss, 2015). Given the quite recent emergence of OTAs, it might be difficult to establish an image as a trusted brand. In the beginning, OTAs should therefore focus on a somewhat more holistic approach, such as comprehensive booking guides. Building on to this, in a study

conducted in 2007, it was found that customers weigh security, in terms of payment and privacy, as the second most important factor when purchasing travel online, after low prices (Jin Kim, Gon Kim, Soo Han, 2007). Though this study is not the most recent one, online security is still a hot topic. Consequently, for an OTA, it is not only important to be able to provide security to its customers, but also to assure them of it, in order to create and maintain long term relations.

7.2.2 Consumer behaviour

Kotler and Armstrong point out four major attributes that affect consumer behaviour, namely cultural, social, personal and psychological. To be able to produce valuable and relevant marketing, it is important that the retailer understands the way customers act and what affects their decisions. It is indisputable that the characteristics about the actual product play a big role in this purchase decision making, but to attain a more universal model, these personal attributes that might also affect the decision making should be studied.

Most people belong to a certain culture. The perhaps most basic form of culture is the one that we grow up with. Individuals will learn and be influenced by the people around them. Adding on to this, there are subcultures, where some of the most well defined ones are either ethnic or age-related (Kotler, Armstrong 2012, p. 136). One of the reasons why it is convenient to identify cultures and subcultures as target groups is because of the shared values, perceptions and ambitions among the members of each group. Marketing towards segments based on culture assumes that there are similarities in the way their communication is regarded. In a research from 2016, on online consumer behaviour, it has been found that there are in fact differences in behaviour between different nationalities (Richard, Reza Habibi, 2016). It is also shown that the differences can be correlated to the general image of the respective countries' cultures. For OTAs this is directly applicable to their core business. For larger OTAs, operating on a global basis, this can be exploited both in terms of promoting popular destinations, but also the way that the website and booking process are set up, as it is possible that people from different nationalities would proceed to the actual booking based on different preferences. If a regression were run on observations separated by the country of the people searching for a flight, different variables might be found to be of significance to different countries. To leverage from this, the results could be used in marketing.

Going back to the purchase decision process, this is what actually motivates customers to purchase a specific product or service. The authors essentially look at the process from two different perspectives. The first one as to the type of purchase, such as habitual or complex, and the second one as what influences the consumer (Kotler, Armstrong 2012, p, 150 - 154). What differs between the categories of the type of purchase is the level of engagement by the consumer. Whereas this level of engagement would increase as the level of complexity of the product increases, it is also a behavior that is seen in overall online purchasing (Kotler, Armstrong 2012, p. 151). According to a study performed by McKinsey & Company, consumers conducting online research, when looking to purchase products of a certain category online, increased from 41% to 50% between 2010 and 2012 (McKinsey & Company 2012). For OTAs, whose core

feature to its products, travels that is, is the search functions allowing customers to compare different flights, this is a real selling point. In a way, OTAs have already incorporated the possibilities for its customers to research the product and its options.

7.3 The marketplace

7.3.1 The business model of an online travel agency

The largest income, accounting for a little more than half of the total, for an OTA are the service fees charged to the customers and mark-ups, being the commission on the price per sold ticket (Hermes Management Consulting S.A., 2010). As they are not dependent on the actual prices, and do not in particular profit from higher prices of the tickets, their marketing model differs from that of an airline company. Likewise is the fact that they only operate online an aspect that would differentiate their marketing model.

7.3.2 The four P's of marketing in an online environment

As to map out how the company's marketing strategy should look, it is convenient to apply the classic model of *The Four P's of Marketing*. The model is about building a marketing strategy around the *product*, which is the goods or services being sold, the *price* it is being sold at, the *place* for the sale, in terms of distribution and finally, *promotion* as in the actual marketing, which are the direct ways that a company can affect its customers (Kotler, Armstrong 2012, p. 51 - 53).

Product

Furthermore, Kotler and Armstrong mean that the product which a company sells, also includes the services provided along with the products. The tangible product that an OTA sells is a flight, which may be paired with other features such as accommodation or car rental. The more intangible services offered along with this, could be divided into two categories, functional and hedonic, according to a study performed by Bernardo, Marimon and del Mar Alonso-Almeida in 2012. The differentiation is based on the perceived value that the customer experiences, where the functional qualities refer to the technical aspects providing the basic features of the websites, such as efficiency, system availability and privacy, whereas the hedonic quality refers to the enjoyment and pleasure of the customer. Despite the functional quality was found to be more important than the hedonic quality in terms of generating a perceived value (Bernardo, Marimon, del Mar Alonso-Almeida, 2012), the additional value from the hedonic qualities are considerable for OTAs to evolve a competitive edge and to maintain the loyalty of their customers. Since OTAs are essentially retail dealers, they are operating with a core product that they cannot change. Consequently, in order to use the product to influence its customers, OTAs should extend the vision of their core-product, from only including the actual flight, to also include the services that they provide. This will enable the OTAs to potentially gain market shares, by providing the customers with the service that they require and appreciate.

Price

Much like any other product sold through a distributor, the price of the airline tickets sold through an OTA consists of the wholesale price, a mark-up and possibly any additional fees. Asides from negotiating the wholesale price with the airline companies, on OTA may of course lower its margins. As they operate in an environment allowing, actually focusing on allowing, the customers to search for and compare the prices for different flights, it is quite safe to say that the price is an actual competitive edge for OTAs. However, OTAs also have an indirect influence on the prices of airline tickets. It has been found that price dispersion in airline tickets increased where there were OTAs present (Roma, Zambuto, Perrone 2014). By displaying prices, additional fees and products, the competition among airline companies, and OTAs, become evident, which has an impact on the prices. Using the online search engines, as OTAs, the market for airline tickets become more distinct and comprehensible for the customers, which may have had a real impact on airline companies' pricing strategies. In a sense, just by merely existing, OTAs provide lower prices for their customers. Then it is up to the OTAs themselves to actively sweetening their prices, by looking to their mark-ups and fees.

Place

Whereas the actual place for the product display and sale are rather definite for an OTA, it might be interesting examining why an airline company would chose to be featured by an OTA, looking towards the other direction of the supply chain. In a study from 2011 it was asserted that the size and the loyalty of the customer base of an airline company was crucial whether the company would sell its tickets through an OTA, in addition to its own website (Koo, Mantin, O'Connor 2011). If the customer base was already of a substantial size, with loyal customers, and especially if the OTA was deemed to be highly competitive, the airline companies were less likely to enter into those types of environments. This aligns with the business model of an OTA, seeing as the idea to provide the customers with comparisons of different flights would not perhaps be of any great interest to customers who already have a preference for specific airline companies. Thus the OTAs would probably have to focus on another key attribute, as the possibility in attracting new customers. Once again, going back to the place of the product display and sale, an OTA is likely to reach a wider range of customers as they, per definition, provide a wider range of services than an airline company.

Promotion

As described by Kotler and Armstrong, direct marketing, and therein online marketing, is a commonly used proxy for a full marketing strategy. The direct marketing, as the name implies, is about connecting directly with the customer, in a form of personalized marketing. The click-only companies are referred to as those companies that only operate in an online environment, such as an OTA. Just as there are differences between online and offline customers, where the online customers are described as proactive, and the offline customers as some-

what passive, the same differences can be distinguished within online versus offline marketing (Kotler, Armstrong 2012, p. 510). As the basic principles of an OTA already assumes an interactive layout, OTAs should focus on online marketing, as to consolidate their active customers. How to further implement different marketing strategies, is discussed throughout this section.

7.4 Communication

7.4.1 Creating trust

CRM, or customer relationship management, incorporates the collection and analysing of customer data, such that it can be used for personalized marketing (Kotler, Armstrong 2012, p. 119 - 120). The purpose is to build solid relationships with the customers, based on the perceived value that the company is able to communicate. As discussed earlier, the collection of information, in order to create personalized marketing is an important factor for OTAs. However, as the perceived trust and sense of security is an important factor for consumers choosing between different retailers (Pappas, 2016), the relation aspect is essential for OTAs. Many consumers tend to value the recommendations of friends and family (Kotler, Armstrong 2012, p. 139), and recommendations are often spread by word-of-mouth. A different angle to this is the online aspect. In an interview by The Boston Consulting Group, the global head of travel strategy for Facebook, Lee McCabe, talked about the many aspects of the relations people have online, particularly on Facebook, and the effects that this has on people's will to purchase travel online (The Boston Consulting Group 2016). The global network, instant communication and sharing of information, would speed up the word-of-mouth process, and widen the effects of it. Thus for an OTA, to create trust and build relations with its customers, could in turn generate new customers, by an online word-of-mouth chain.

7.4.2 Channels

On the topic of marketing channels, these are described as a chain of intermediaries, which together create and distribute the product, from manufacturing to the purchase by the final customer (Kotler, Armstrong 2012, p. 344). The parties in these channels are dependent on each other, as the product in a sense passes between levels in the channel, from the producer, to the wholesaler, to the retailer and so on. However, for online retailers such as OTAs selling a type of services, there may be more layers to this approach. It is evident that online marketing is an efficient tool for OTAs, accounting for their actual business being online, and thus the logistic aspects to that. Specifically, the usage of social media in marketing purposes, might be an attractive alternative. In terms of building on to the relations, as previously discussed, social media offers a two way communication between the company and the customers, simultaneously as providing the classical marketing aspects, in communicating a brand and its products. As customers engage with the company on social media, by following, liking or for example sharing company news or features, the company increases its brand health (Syakirah Ahmad, Musa, Harris Mior Harun, 2016). This phenomenon can be interpreted as a cycle, where proactive marketing on social media encourages customers to interact, and the brand is strengthened. A

strong brand evidently goes hand in hand with building relations, an important aspect for OTAs.

7.5 Conclusion

There are certainly several factors that seem to be valuable when analyzing the potential layout of a marketing model of an OTA. Some key attributes, such as the online environment, or the products the OTAs sell need to be taken into account. One of the aspects that permeates the business model of OTAs, is the trust and relations that are built, and hopefully maintained, with the customers. As the usage of customer data could in this sense be used in terms of personalized marketing, it is crucial that customers are ensured that their information is handled properly. The online environment has generated a new and more complex customer behaviour and habits, which allows for new types of marketing. This has had an impact on the classical interpretation and application of the model of the four P's, where services and a wide range of products offered, play a great part in both maintaining customers, but also to attract new, and to reach out to new platforms. The online business model and the impact of social media has made its mark on the way that the brand is perceived, which the OTAs could leverage from.

The key take away, is that the online business has evolved both more sophisticated customers, but also a recent requirement for security. The customers require new and more specialized marketing, whereas security is communicated and maintained through a solid relationship, that can be built using the new online channels of marketing.

8 References

Litterature

Lang H., 2015 *Elements of Regression Analysis*, Harald Lang

Kennedy P., 2008, *A Guide to Econometrics*, 6th edition, Wiley-Blackwell, Malden, ISBN 978-1-4051-8258-4

Kotler P., Armstrong G., 2012, *Principles of Marketing*, 14th edition, Pearson Education Inc., New Jersey, ISBN 10: 0-13-216712-3, ISBN 13: 978-0-13-216712-3

Ferrell O. C., Hartline M. D., 2011, *Marketing Management Strategies*, International Edition, Cengage Learning, ISBN-13: 978-0-538-46744-5, ISBN-10: 0-538-46744-4

Meerman Scott D., 2008, *The New Rules of Viral Marketing*, David Meerman Scott

Journals

Roger-Monzó V., Martí-Sánchez M., Guijarro-García M., 2015, *Using online consumer loyalty to gain competitive advantage in travel agencies*, Journal of Business Research, vol. 68, issue 7, p. 1638-1640

Turner P. A., Hoon Lim S., 2015, *Hedging jet fuel price risk: The case of U.S. passenger airlines*, Journal of Transport Management, vol. 44-45, p. 54-64

Fox J., Monette G., 1992, *Generalized Collinearity Diagnostics*, Journal of the American Statistical Association, vol. 87, issue 417, p. 178-183

Bleier A., Eisenbeiss M., 2015, *The Importance of Trust for Personalized Online Advertising*, Journal of Retailing, vol. 91, issue 3, p. 390-409

Jin Kim D., Gon Kim W., Soo Han J., 2007, *A perceptual mapping of online travel agencies and preference attributes*, Tourism Management, vol. 28, issue 2, p. 591-603

Richard M. O., Reza Habibi M., 2016, *Advanced modeling of online consumer behaviour: The moderating roles of hedonism and culture*, Journal of Business Research, vol. 69, issue 3, p. 1103-1119

Bernardo M., Marimon F., del Mar Alonso-Almeida M., 2012, *Functional quality and hedonic quality: A study of the dimension of e-service quality of online travel agencies*, Information Management, vol. 49, issues 7-8, p. 342-347

Roma P., Zambuto F., Perrone G., 2014, *Price dispersion, competition, and the role of online travel agents: Evidence from business routes in the Italian airline market*, Transportation Research Part E: Logistics and Transportation Review, vol. 69, p. 146-159

Koo B., Mantin B., O'Connor P., 2011, *Online distribution of airline tickets: Should airlines adopt a single or a multi-channel approach?*, Tourism Management, vol. 31, issue 1, p. 69-74

Pappas N., 2016, *Marketing strategies, perceived risks, and consumer trust in online buying behaviour*, Journal of Retailing and Consumer Services, vol. 29, p. 92-103

Syakirah Ahmad N., Musa R., Harris Mior Harun M., 2016, *The Impact of Social Media Content Marketing (SMCM) towards Brand Health*, Procedia Economics and Finance, vol. 37, p. 331-336

Reports

Hellström D., 2013, *Flygtendenser - Statistik, analys och information från Transportstyrelsen*, Transportstyrelsen, p. 5

2016, *Tourism statistics*, Eurostat, ISSN 2443-8219

Hazan E., Wagener N., 2013, *iConsumers: Life online*, (Telecommunications, Media and Technology Practice), McKinsey & Company, p. 10

2010, *Understanding Online Travel Agencies' Cost Drivers and Ways to Optimize Business in Europe*, Hermes Management Consulting S.A., commissioned by Amadeus IT Group SA

Guggenheim J., 2016, *Facebook on the Future of Travel*, The Boston Consulting Group

9 Appendix

9.1 Long distant destinations

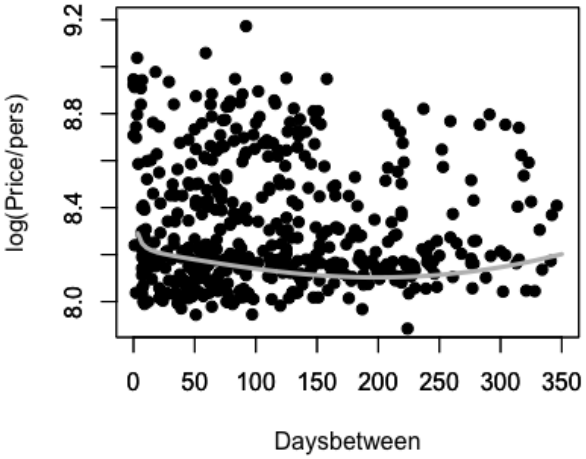

Figure 9.1: Scatter plot NYC

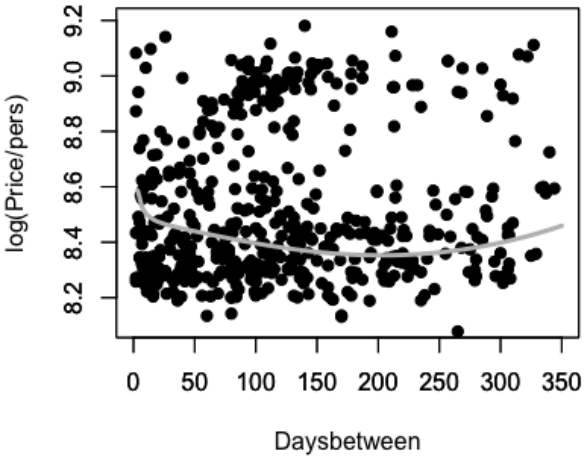

Figure 9.2: Scatter plot LAX

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	8.42	0.01	0.89	0.00	8.41	8.44
depweekdMon	-0.05	0.00	0.00	0.00	-0.05	-0.05
depweekdTues	-0.04	0.00	0.00	0.00	-0.05	-0.04
depweekdWed	-0.08	0.00	0.01	0.00	-0.08	-0.07
depweekdThurs	-0.04	0.00	0.00	0.00	-0.04	-0.04
depweekdFri	-0.08	0.00	0.01	0.00	-0.08	-0.07
depweekdSat	-0.03	0.00	0.00	0.00	-0.04	-0.03
homeweekdMon	-0.02	0.00	0.00	0.00	-0.02	-0.02
homeweekdTues	0.01	0.00	0.00	0.00	0.00	0.01
homeweekdWed	-0.01	0.00	0.00	0.00	-0.01	-0.00
homeweekdThurs	0.03	0.00	0.00	0.00	0.03	0.04
homeweekdFri	0.05	0.00	0.00	0.00	0.04	0.05
homeweekdSat	0.12	0.00	0.03	0.00	0.12	0.13
fromBMA	-0.02	0.00	0.00	0.00	-0.02	-0.01
fromCPH	0.04	0.00	0.01	0.00	0.04	0.04
fromGOT	-0.00	0.00	0.00	0.25	-0.00	0.00
fromNYO	0.68	0.03	0.00	0.00	0.62	0.74
depmonth(1,2]	-0.15	0.00	0.02	0.00	-0.16	-0.14
depmonth(2,3]	0.02	0.00	0.00	0.00	0.01	0.02
depmonth(3,4]	0.05	0.00	0.00	0.00	0.05	0.06
depmonth(4,5]	0.04	0.00	0.00	0.00	0.04	0.05
depmonth(5,6]	0.25	0.00	0.06	0.00	0.25	0.26
depmonth(6,7]	0.57	0.00	0.22	0.00	0.56	0.57
depmonth(7,8]	0.30	0.00	0.05	0.00	0.29	0.31
depmonth(8,9]	0.07	0.00	0.00	0.00	0.07	0.08
depmonth(9,10]	0.05	0.00	0.00	0.00	0.05	0.06
depmonth(10,11]	-0.08	0.00	0.00	0.00	-0.09	-0.08
depmonth(11,12]	0.18	0.00	0.03	0.00	0.17	0.19
daysbetween	-0.00	0.00	0.00	0.00	-0.00	-0.00
squaredays	0.00	0.00	0.01	0.00	0.00	0.00
logdays	-0.13	0.01	0.00	0.00	-0.14	-0.12
I(logdays^2)	0.03	0.00	0.00	0.00	0.02	0.03
holiday	0.15	0.00	0.08	0.00	0.15	0.15

Table 9.1: Final result NYC

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	8.75	0.01	0.89	0.00	8.73	8.78
depweekdMon	-0.04	0.00	0.00	0.00	-0.05	-0.04
depweekdTues	-0.06	0.00	0.01	0.00	-0.06	-0.05
depweekdWed	-0.06	0.00	0.01	0.00	-0.07	-0.06
depweekdThurs	-0.04	0.00	0.00	0.00	-0.04	-0.03
depweekdFri	-0.03	0.00	0.00	0.00	-0.04	-0.03
depweekdSat	-0.02	0.00	0.00	0.00	-0.02	-0.02
homeweekdMon	-0.02	0.00	0.00	0.00	-0.02	-0.02
homeweekdTues	-0.03	0.00	0.00	0.00	-0.03	-0.02
homeweekdWed	-0.02	0.00	0.00	0.00	-0.02	-0.01
homeweekdThurs	-0.00	0.00	0.00	0.36	-0.01	0.00
homeweekdFri	0.04	0.00	0.00	0.00	0.03	0.04
homeweekdSat	0.02	0.00	0.00	0.00	0.01	0.02
fromBMA	0.06	0.01	0.00	0.00	0.04	0.07
fromCPH	0.04	0.00	0.00	0.00	0.03	0.04
fromGOT	-0.03	0.00	0.00	0.00	-0.03	-0.03
depmonth(1,2]	-0.15	0.00	0.02	0.00	-0.16	-0.15
depmonth(2,3]	-0.03	0.00	0.00	0.00	-0.04	-0.03
depmonth(3,4]	-0.04	0.00	0.00	0.00	-0.05	-0.04
depmonth(4,5]	-0.05	0.00	0.00	0.00	-0.05	-0.04
depmonth(5,6]	0.17	0.00	0.03	0.00	0.16	0.17
depmonth(6,7]	0.55	0.00	0.22	0.00	0.54	0.55
depmonth(7,8]	0.31	0.00	0.06	0.00	0.30	0.32
depmonth(8,9]	-0.03	0.00	0.00	0.00	-0.03	-0.02
depmonth(9,10]	-0.06	0.00	0.00	0.00	-0.07	-0.05
depmonth(10,11]	-0.12	0.00	0.01	0.00	-0.12	-0.11
depmonth(11,12]	0.13	0.00	0.02	0.00	0.13	0.14
daysbetween	-0.00	0.00	0.00	0.00	-0.00	-0.00
squaredays	0.00	0.00	0.01	0.00	0.00	0.00
logdays	-0.17	0.01	0.00	0.00	-0.19	-0.15
I(logdays^2)	0.03	0.00	0.00	0.00	0.03	0.04
holiday	0.11	0.00	0.03	0.00	0.10	0.11

Table 9.2: Final result LAX

Weekday	Difference (%)	Weekday	Difference (%)
Wednesday	-7	Tuesday	-6
Friday	-7	Wednesday	-6
Monday	-5	Monday	-4
Tuesday	-4	Thursday	-4
Thursday	-4	Friday	-3
Saturday	-3	Saturday	-2

Table 9.3: Weekday of the outbound flight for NYC and LAX, Sunday as benchmark

Weekday	Difference (%)	Weekday	Difference (%)
Monday	-2	Tuesday	-3
Wednesday	-1	Monday	-2
Tuesday	1	Wednesday	-2
Thursday	4	Thursday	0
Friday	5	Saturday	2
Saturday	13	Friday	4

Table 9.4: Weekday of the inbound flight for NYC and LAX, Sunday as benchmark

Airport	Difference (%)	Airport	Difference (%)
BMA	-2	GOT	-3
GOT	0	CPH	4
CPH	4	BMA	6
NYO	97	NYO	-

Table 9.5: Departure airport for NYC and LAX, ARN as benchmark

Month	Difference (%)	Month	Difference (%)
February	-14	February	-14
November	-8	November	-11
March	2	October	-6
May	5	May	-5
October	5	April	-4
April	6	March	-3
September	7	September	-3
December	20	December	14
June	28	June	18
August	35	August	36
July	77	July	73

Table 9.6: Month of departure for NYC and LAX, January as benchmark

	Number of days		Number of days
lowest price	199	lowest price	203

Table 9.7: The number of days between booking and the outbound flight for NYC and LAX

9.2 European destinations

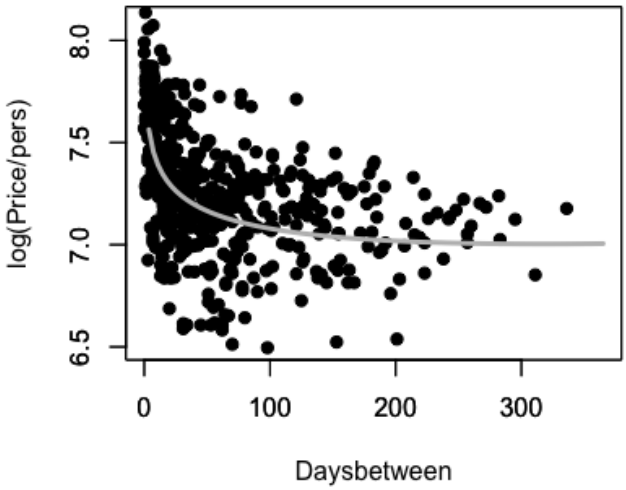

Figure 9.3: Scatter plot PAR

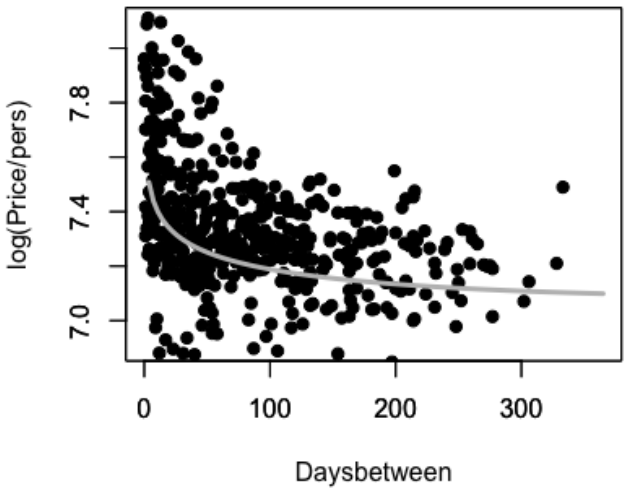

Figure 9.4: Scatter plot ROM

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	7.77	0.01	0.97	0.00	7.76	7.78
depweekdMon	-0.05	0.00	0.00	0.00	-0.06	-0.05
depweekdTues	-0.08	0.00	0.01	0.00	-0.08	-0.07
depweekdWed	-0.03	0.00	0.00	0.00	-0.04	-0.03
depweekdThurs	0.07	0.00	0.01	0.00	0.07	0.08
depweekdFri	0.05	0.00	0.00	0.00	0.05	0.06
depweekdSat	0.00	0.00	0.00	0.53	-0.00	0.01
homeweekdMon	-0.10	0.00	0.02	0.00	-0.11	-0.10
homeweekdTues	-0.19	0.00	0.06	0.00	-0.19	-0.18
homeweekdWed	-0.21	0.00	0.07	0.00	-0.21	-0.20
homeweekdThurs	-0.17	0.00	0.04	0.00	-0.17	-0.16
homeweekdFri	-0.13	0.00	0.03	0.00	-0.14	-0.13
homeweekdSat	-0.07	0.00	0.01	0.00	-0.08	-0.07
fromBMA	-0.00	0.01	0.00	0.79	-0.02	0.01
fromCPH	-0.07	0.00	0.01	0.00	-0.07	-0.06
fromGOT	0.06	0.00	0.01	0.00	0.05	0.06
fromNYO	-0.30	0.00	0.16	0.00	-0.31	-0.30
depmonth(1,2]	-0.23	0.01	0.04	0.00	-0.24	-0.22
depmonth(2,3]	-0.03	0.00	0.00	0.00	-0.04	-0.02
depmonth(3,4]	0.11	0.00	0.01	0.00	0.10	0.12
depmonth(4,5]	0.07	0.00	0.01	0.00	0.07	0.08
depmonth(5,6]	0.17	0.00	0.02	0.00	0.16	0.18
depmonth(6,7]	0.29	0.00	0.05	0.00	0.28	0.30
depmonth(7,8]	0.12	0.01	0.01	0.00	0.10	0.13
depmonth(8,9]	0.06	0.00	0.00	0.00	0.05	0.07
depmonth(9,10]	0.03	0.00	0.00	0.00	0.02	0.04
depmonth(10,11]	-0.05	0.00	0.00	0.00	-0.06	-0.04
depmonth(11,12]	0.21	0.00	0.04	0.00	0.20	0.22
daysbetween	0.00	0.00	0.01	0.00	0.00	0.00
logdays	-0.16	0.00	0.19	0.00	-0.16	-0.16
holiday	0.22	0.00	0.15	0.00	0.22	0.22

Table 9.8: Final result PAR

	Estimate	Std.Error	Eta.sq	p.value	lower	upper
(Intercept)	7.64	0.01	0.97	0.00	7.63	7.65
depweekdMon	-0.00	0.00	0.00	0.75	-0.01	0.01
depweekdTues	-0.02	0.00	0.00	0.00	-0.03	-0.01
depweekdWed	0.01	0.00	0.00	0.00	0.00	0.02
depweekdThurs	0.07	0.00	0.01	0.00	0.06	0.07
depweekdFri	0.03	0.00	0.00	0.00	0.03	0.04
depweekdSat	0.04	0.00	0.00	0.00	0.03	0.04
homeweekdMon	-0.07	0.00	0.02	0.00	-0.08	-0.07
homeweekdTues	-0.12	0.00	0.04	0.00	-0.13	-0.12
homeweekdWed	-0.15	0.00	0.04	0.00	-0.15	-0.14
homeweekdThurs	-0.14	0.00	0.04	0.00	-0.14	-0.13
homeweekdFri	-0.10	0.00	0.02	0.00	-0.11	-0.10
homeweekdSat	-0.03	0.00	0.00	0.00	-0.04	-0.03
fromBMA	0.01	0.01	0.00	0.03	0.00	0.02
fromCPH	-0.04	0.00	0.01	0.00	-0.04	-0.04
fromGOT	0.08	0.00	0.03	0.00	0.08	0.09
fromNYO	-0.20	0.00	0.07	0.00	-0.19	-0.18
depmonth(1,2]	-0.13	0.01	0.01	0.00	-0.14	-0.12
depmonth(2,3]	0.03	0.00	0.00	0.00	0.02	0.04
depmonth(3,4]	0.08	0.00	0.01	0.00	0.08	0.09
depmonth(4,5]	0.03	0.00	0.00	0.00	0.03	0.04
depmonth(5,6]	0.09	0.00	0.01	0.00	0.08	0.10
depmonth(6,7]	0.23	0.00	0.03	0.00	0.21	0.23
depmonth(7,8]	0.11	0.01	0.01	0.00	0.09	0.11
depmonth(8,9]	0.09	0.01	0.01	0.00	0.07	0.09
depmonth(9,10]	0.07	0.00	0.01	0.00	0.07	0.08
depmonth(10,11]	-0.03	0.00	0.00	0.00	-0.04	-0.02
depmonth(11,12]	0.22	0.01	0.04	0.00	0.21	0.23
daysbetween	0.00	0.00	0.00	0.00	0.00	0.00
logdays	-0.10	0.00	0.10	0.00	-0.10	-0.10
holiday	0.18	0.00	0.13	0.00	0.18	0.19

Table 9.9: Final result ROM

Weekday	Difference (%)	Weekday	Difference (%)
Tuesday	-7	Tuesday	-2
Monday	-5	Monday	0
Wednesday	-3	Wednesday	1
Saturday	0	Saturday	4
Friday	6	Friday	4
Thursday	8	Thursday	7

Table 9.10: Weekday of the outbound flight for PAR and ROM, Sunday as benchmark

Weekday	Difference (%)	Weekday	Difference (%)
Wednesday	-19	Wednesday	-14
Tuesday	-17	Thursday	-13
Thursday	-15	Tuesday	-12
Friday	-12	Friday	-10
Monday	-10	Monday	-7
Saturday	-7	Saturday	-3

Table 9.11: Weekday of the inbound flight for PAR and ROM, Sunday as benchmark

Airport	Difference (%)	Airport	Difference (%)
NYO	-26	NYO	-18
CPH	-6	CPH	-4
BMA	0	BMA	1
GOT	6	GOT	8

Table 9.12: Departure airport for PAR and ROM, ARN as benchmark

Month	Difference (%)	Month	Difference (%)
February	-21	February	-12
November	-5	November	-3
March	-3	March	3
October	3	May	3
September	6	October	8
May	8	April	9
April	12	September	9
August	12	June	10
June	19	August	11
December	23	December	25
July	34	July	26

Table 9.13: Month of departure for PAR and ROM, January as benchmark

	Number of days		Number of days
lowest price	316	lowest price	685

Table 9.14: The number of days between booking and the outbound flight for PAR and ROM

TRITA -MAT-K 2016:18
ISRN -KTH/MAT/K--16/18--SE