|  |  |
| --- | --- |
| Afbeeldingsresultaat voor han nieuwe logo | **DATA DRIVEN DESISCION MAKING IN BUSINESS**  Course: individual researchproject  Stan van Gaal (Studentnummer: 634716)  Versie 1.0  Supervisor: John Smit  Minor: DDDDMIB  22 december 2022  Arnhem |
|  |  |

Inhoud

[1. Preface 3](#_Toc123734875)

[2. Introduction 4](#_Toc123734876)

[3. Literature review 5](#_Toc123734877)

[3.1 Process & CRISP 5](#_Toc123734878)

[4. Goal 6](#_Toc123734879)

[5. Research 7](#_Toc123734880)

[5.1 What are important characteristics to asses the price of a flight? 7](#_Toc123734881)

[5.1.2 sub conclusion 8](#_Toc123734882)

[5.2 What are the relevant datafeatures in this domain? 9](#_Toc123734883)

[5.2.1 Comparing features with literature 10](#_Toc123734884)

[5.2.2 sub conclusion 11](#_Toc123734885)

[5.3 How can the data be modeled to predict the price of a flight? 12](#_Toc123734886)

[5.3.1 Creating dummies 12](#_Toc123734887)

[5.3.2 Checking missing values 13](#_Toc123734888)

[5.3.3 Multicollinearities and scatterplot matrix 13](#_Toc123734889)

[5.3.4 Outliers 14](#_Toc123734890)

[7.4.5 Modeling 15](#_Toc123734891)

[5.4 How is the product available for the end-user? 16](#_Toc123734892)

[6. Data science Techniques and Tools 17](#_Toc123734893)

[7. Methodology 18](#_Toc123734894)

[8. Conclusion 19](#_Toc123734895)

[9. Reflection 20](#_Toc123734896)

[11. Bibliographie 21](#_Toc123734897)

[12. Appendix 22](#_Toc123734898)

# 1. Preface

My name is Stan van Gaal. I am currently studying on the HAN in Arnhem. During my bachelor’s degree, there is a minor in planning. I am following the minor Data Driven Desiscion Making in Business for five months. The time period of this research is November 2022 untill January 2023. The goal of this project is to demonstrate that I can independently apply the material learned within this minor to a case from the internet. From the minor, a project report is included in which I take a closer look at what has been delivered.

I chose this project because I find it interesting to find patterns in the data and then attach steering information to this. You also really have a working end product in which you can predict a price per flight based on variables. For example, you could recommend to others which flights are cheap to book. That sounds really cool.

Inside the HAN, I would like to express our gratitude to my projectmanager Tijmen Weber and John Smits for the guidance I were able to get during this project.

This report is called: ‘Inidividual researchproject’

Lots of reading pleasure,

Stan van Gaal

Arnhem, 09-01-2023

# 2. Introduction

This document is an individual project report that provides a detailed description of the work done (quantity and quality). Furthermore, the development of my professional expertise during the project is clearly shown. The purpose of this project report is therefor to justify my own contribution and to provide insight into my learning process.

What I want to achieve with this project report is to map out my learning process and explain the apllied techniques. Ultimately, the theory is applied in practice. This in order to be able to improve myself in the field of data science skills from here. My challange mainly will be in modeling within R. As a business-IT background, this is a bit more technical and complicated for me, which I see as a challange.

I started searching on Kaggle for a dataset to work with. Here I came across countless datasets. It was therefore difficult for me to find one. In the end i opted for a ‘flight prediction’ dataset. The reason why I chose this project specific is because it is for everyone to imagine what this dataset looks like. It is easy to use and you already have an idea of the content of the variables in advance. Of course everyone has an idea of what a flight looks like. This ensured that the dataset was easy to understand and that I could get started quickly. During another course within the minor, the dataset was very complicated, which meant that the data understandingpart took a lot of time. I wanted to save myself this during the individual project. That is why I chose an easily understandable dataset.

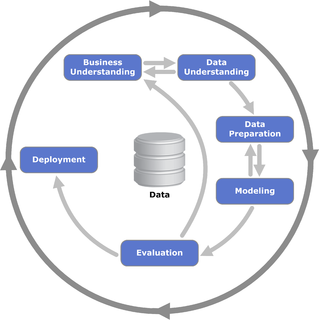
The report is written by Stan van Gaal, student at the HAN in Arnhem and is written for supervisor Tijmen Weber and John Smit from the ‘Hogeschool Arnhem en Nijmegen’. The report is divided into chapters, and it becomes evident what the development's structure should look like.

# 3. Literature review

In this chapter the literature and models for this project will be clarified. For this project it is important to use and follow the CRISP model.

## 3.1 Process & CRISP

“CRISP” stands for “Cross Industry Standard Process.” The model serves as a framework for planning and managing data mining projects. The CRISP steps form an iterative process in which there is constant movement between the different process steps. The model is dynamic and in practice is often combined with other agile and lean projectmanagement approaches. The six process steps are explained below. The model is showed in *Figure 1*.



*Figure 1: CRISP Model*

So how will this model be used within my project, starting with:

1. **Business understanding:** To get a good understanding of the project a broader business understanding is needed. Therefor I need knowledge about the main goals and achievements from the business. But in my case, it isn’t a real business case but a dataset that can be used to predict flight prices on different variables.
2. **Data understanding**: After the project scope is defined, the data will be examined to see which steps need to be taken before modeling can be done at all. I will look within the dataset at the steps that I received from the course modeling. Think of outliers, NA values, standardization etc.
3. **Data preparation**: In this phase the data will prepared for modeling. So, the above steps will be elaborated within R.
4. **Modeling**: The next step after data preparation is modeling. Regression models will be made with which you can predict how expensive a flight is bases on variables. Also, other research questions that have been prepared in advance will also be answered.
5. **Evaluation**: In this step, the product resulting from the modeling step is compared with the business understanding.
6. **Deployment**: In this phase the product will be made available to the user. In this case it will be uploaded on GitHub so everyone can look into the details of the project.

# 4. Goal

The project goal involves a number of goals. I look at which variables within the dataset have the most impact on the price. This so that people know what to look out for when buying a plane ticket. A model will also be built that can predict how expensive an airline ticket will be bases on variables used in the dataset. However, this is limited to data from the dataset. It is also a goal to justify my own choices within this project. It will be explained how the theory of school as finally been applied within this project.

This project also helps me in my personal development in the data science. I chose this minor to eventually gain some more knowledge about data science for a possible follow-up stop to my bachelor’s HBO-ICT BIM.

# 5. Research; ‘What ensures that you can book a plane ticket as cheaply as possible?’

## 5.1 What are important characteristics to asses the price of a flight?

In this chapter, a literature study is used to investigate which variables make a flight so expensive. This is done bases on internet research. Furthermore, in this project I am of course also bound by the variables that are used in the dataset. This prevents me from adding anything. However, this will provide insight into whether this dataset provides the correct information to determine how expensive a flight is. Regardless of the results of this, I will always continue the investigation. Various factors that influence a ticket price will be described below. The next sub-question examines which factor can be found within the dataset.

**Business understanding**

**Dinstance**

In first place the distance is a important factor to determining the price of a flight ticket. Depending on travel time, a ticket will often become more expensive. Although this is almost related to the ticketprice, it does not always have to be a decisive factor. There are of ocourse also important other factors that are described below.

**Market forces**

It is important here how popular the ticket is at what time, or suply and demand. As a result, it can cost a lot if, for example, it is the last chair in the airplane. While at other times it is a lot cheaper. Prices are constantly changing as reseravtions fill up or seats become avaible again. However, this factorwill not be included in this dataset. A price for an airline ticket has been calculated per flight.

**Peak season**

Another important factor is seasonality. This means that when certain areas are in demand for a certain amount of time, the airline ticket will see a spike in price. For example, think of tourist attractions in the summer. Since demand is high, ticket prices will therefore rise.

**Flight timing**

The time of a flight can also lead to a price difference in the airline tickets. Think of early morning flights or night flights. The data of booking also affects the ticket price. If you book months in advance, it will cost less than buying a plane ticket 1 or 2 weeks in advance. Although booking last minute can also yield extra benefits. Timing therefore plays an important role in receiving the best ticket prices.

**Flight travel type**

Type of flight trip is the trip you indulge in after purchasing the ticket. Think of direct flights with shorter travel time or you have an indirect flight with several stopovers and longer travel time. Stopovers also influence the ticket price.

The length of a stay also often influences the price of a flight. If you only stay for a few days (2 to 3 days) at a destination, the airline’s assume a business trip, which will cause the price of a ticket to rise.

**Competition**

Currently, there is no monopoly within the airline industry, so that means that competition is also key when it comes to flight ticket prices fluctuations. So airlines will come up with offers and discounts to help their business and win form the competition.

**Oil prices**

Currently you can not escape from the rising oil prices because of the ukrian war. The prices of fuel is high and going up. Also airlines need to adjust prices to account for that expense. Fuel is one of the biggest costs of operating a flight.

(Intermiles, 2020)

In the next chapter we will look at the different datafeatures. These will be compared to the above factors. Which can be measured and which cannot be measured within the dataset.

### 5.1.2 sub conclusion

So, consider the sub-question, ‘What are important charactersitics to asses the price of a flight?’. We can conclude that distance and popularity are the main factors that determine the ticket prices. As with any other valuable product, airlines determine the price of an airline ticket based on supply and demand. When demand is low, prcies are lowered to boost sales. Is demand high? Then the air ticket prices are increased to take advantage of more profit. However, there are still several variables that also affect the ticket price. These only have less influence than the two mentioned above. Whether this is ture, we will find out in the modeling part. It can be concluded from this which variables now have the most impact on the price.

## 5.2 What are the relevant datafeatures in this domain?

**Dataset**

The dataset contains information about flight booking options from the website Easemytrip for flight travel between india’s top 6 metro cities. There are 300.361 records and 11 features in the cleaned dataset.

**Data collection and methodology**

Octoparse scraping tool is used to extract data from the website ‘Ease my trip’. It was collected in two parts: one for economy class tickets and another for businessclass tickets. A total of 300.261 distinct flight booking options was extracted from the site. The data is collected for 50 days, from February 11th to march 31st, 2022.

(BATHWA, 2022)

**Data understanding**

**Economy dataset**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Meaning |
| Date | DATE | Date of measurement |
| Airline | TEXT | The name of the airline company (six different airlines) |
| Ch\_code | TEXT | Part of flightcode |
| Num\_code | INT | Part of flightcode |
| Dep\_time | TIME | Takeoff time |
| From | TEXT | Country of departure |
| Time\_taken | TEXT | Flight duration |
| Stop | TEXT | Count of stopovers (3 distinct values) |
| Arr\_time | TIME | Information about the arrival time |
| to | TEXT | City where the flight will land |
| Price | Integer | Information ticket price |

**Business dataset**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Meaning |
| Date | DATE | Date of measurement |
| Airline | TEXT | The name of the airline company (six different airlines) |
| Ch\_code | TEXT | Part of flightcode |
| Num\_code | INT | Part of flightcode |
| Dep\_time | TIME | Takeoff time |
| From | TEXT | Country of departure |
| Time\_taken | TEXT | Flight duration |
| Stop | TEXT | Count of stopovers (3 distinct values) |
| Arr\_time | TIME | Information about the arrival time |
| to | TEXT | City where the flight will land |
| Price | Integer | Information ticket price |

These two datasets were eventually mergid into a ‘clean dataset’. Various adjustments have been added, such as: combining Ch\_code and Num\_code, changing variable names, types. In the future, it will be examined which variables corresond to the literature study to predict a flight price ticket. This is all done by the owner of the dataset. The provided dataset I work with is the clean dataset below.

Looking at the cleaned dataset, the data collected is seprated over 11 variables. These items will be used for the modeling part. However a bit of data preperation will precede this because it is not possible to model with TEXT types. This will be explained in the chapter below.

**Clean dataset**

|  |  |  |
| --- | --- | --- |
| Variable | Type | Meaning |
| Airline | TEXT | The name of the airline company (six different airlines) |
| Flight | TEXT | Flight code |
| Source\_city | TEXT | Country of departure |
| Departure\_time | TEXT | Takeoff time |
| Stops | TEXT | Count of stopovers (3 distinct values) |
| Arrival\_time | TEXT | Arrival time |
| Destination\_city | TEXT | City where the flight will land. It is a categorical feature having 6 unique cities. |
| Class | TEXT | A categorical feature that contains information on seat class; it hass two distinct values: Business and Economy |
| Duration | INT | Flightduration between cities in hours |
| Days\_left | INT | This is a derived attribute calculated by subtracting the travel data from the booking date. |
| Price | INT | Information ticket price |

### 5.2.1 Comparing features with literature

As expected, it is possible to do calculations with all variables within this dataset. Otherwise it was not cleaned. These variables fulfill the following variables from the literature study:

* Distance
* Competition
* Flight travel type
* Flight timing

Unfortunately, the other variables in the literature study are not taken into account. This is partly because this a snapshot of the dataset. This means that this data is from one day (11 February) so it is not possible to do calculations using peak season. Also the influence of oil prices is also outside the scope of this case. However, this the oil price will be the same for every flight since this data is about one specific day.

During the modeling, the difference in the price during the changes of the variables below will be looked at. For example; what happens when a plane has multiple stops instead of a direct flight. We will also look at which variables have the most impact on the price of an airline ticket. From the literature study, this will therefore be the distance (which can be measured in the dataset) and also the market forces and peak season. Unfortuantely, this cannot be measured.

|  |  |  |
| --- | --- | --- |
| Variable | Useful? | Explanation |
| Airline | YES | Checking differency between airline organisation |
| Flight | YES | Some type of flight cheaper than other types? |
| Source\_city | YES | Difference between source city? |
| Departure\_time | YES | Difference between departure time? |
| Stops | YES | Many stops, cheaper flights? |
| Arrival\_time | YES | Difference between arrival time? |
| Destination\_city | YES | Difference between destination city? |
| Class | YES | Difference between classes |
| Duration | YES | Difference between duration |
| Days\_left | YES | Amount of left days |
| Price | YES | ~ |

### 5.2.2 sub conclusion

There are 2 main groups of datasets; Economy- and businessdataset. These two datasets are merged into one dataset called ‘Cleaned dataset’. The variables of ‘Cleaned dataset’ are clear. The meaning of the variables were clearly explained on Kiggle’s site where the owner provides more information about this. The dataset is also not very diffuclt since everyone has an idea about flights.

## 5.3 How can the data be modeled to predict the price of a flight?

First of all, we need to look at what steps need to be taken before the modeling can begin. This will be explained under the datapreparation section. The dataset being worked on is ‘Cleaned dataset’.

The language ‘R’ is used for modeling. This was chosen because most knowledge was gained in this language during the minor Data Driven Desiscion Making in Business. Experience has already been gained with ‘R’ for three courses within the minor. So it seemed natural to also use ‘R’ for this command. In the R script the libraries are loaded. These are required to use certain functions to do the data preperation part.

**Data preperation**

### 5.3.1 Creating dummies

It is important that dummies are made of variables in which a string is used. Only with INT can modeling be done. In this dataset, only a few unique values are included for each variable:

|  |  |
| --- | --- |
| Variable | Unique values |
| Airline | 6 different airlines |
| Flights | Categorical feature |
| Source city | 6 unique cities |
| Departure time | 6 unique time labels |
| Stops | 3 distinct values |
| Arrival time | 6 distinct time labels |
| Destination city | 6 unique cities |
| Class | 2 distinct values |
| Duration | NA |
| Days left | NA |
| Price | NA |

For example; each airline organization gets its own number, each source city gets its own number etc. As explained above, this ensures that modeling can taken place.

To create dummy data the function ‘dummy\_cols’ is used. Within a variable, this function returns a value of 1 when a categorical event occurs and a 0 when it does not. All columns are selected that have a number of (max) unique TEXT values. These are indicated in the table above. No calculation can be performed based on the flight code. This is a unique flight number.

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

**Result:**

**Afbeelding met tekst, wit

Automatisch gegenereerde beschrijving**The table below shows the outcome of the dummy columns. Here you can see that a 1 is returned each time the value meets the variables and a 0 when is does not. In the first row you can see that the airline is ‘Airline SpiceJet’ because it returns a ‘1’.

### 5.3.2 Checking missing values

The dataset is checked for missing values. All values are filled in this dataset. This means that there are no missing values. This is very important to measure, because if the missing values are higherthan 5% of the whole dataset, the modelling cannot be done accurately enough. In our case this is not a problem.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

### 5.3.3 Multicollinearities

**Multicollinearities:**

Multicollinearity occurs when independent variables are ccorrelated in a regression model. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems in fitting the model an dinterpreting the results.

In the image below, only the columns with numeric values are checked. The coefficient is 1 for all three columns. That means it shows a perfect positive correlation, or a direct relationship. So it is a linear relationship.

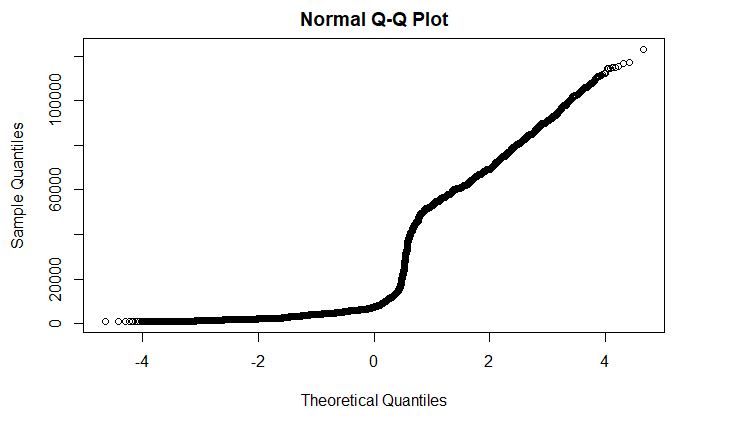
Afbeelding met tekst

Automatisch gegenereerde beschrijving

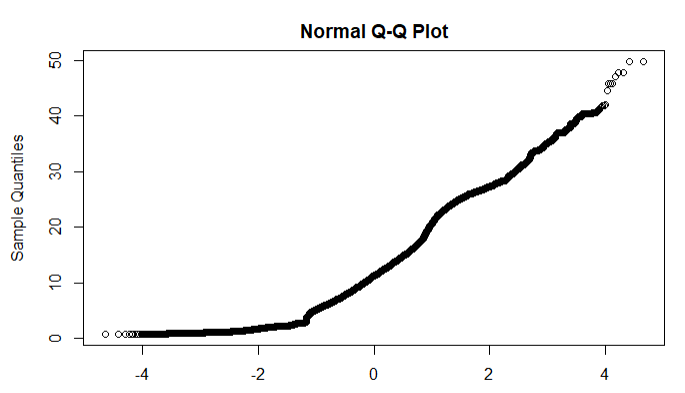
### 5.3.4 Outliers

Removing outliers means that you remove extreme values from your data set before performing any analyses. The goal is to remove all incorrect data, but to keep the real extreme values. To find the outliers in R the qqnorm plot is used. For each variable, the plot is checked to see if an outlier can be found. In the follow examples of the qqnorm plots, the only variables used are price and duration. It is not possible to use the other variables because of the dummy data or unique ID (flight code)

**Price:**

****

**Duration:**

****

Currently there are no worries about the outliers in the examples above. So it is not applicable to remove outliers within these plots.

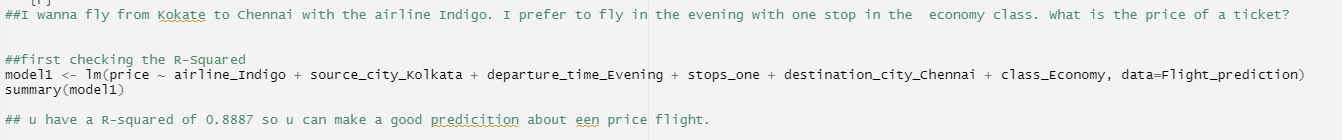
### 5.4.1 Modeling

**Data modeling**

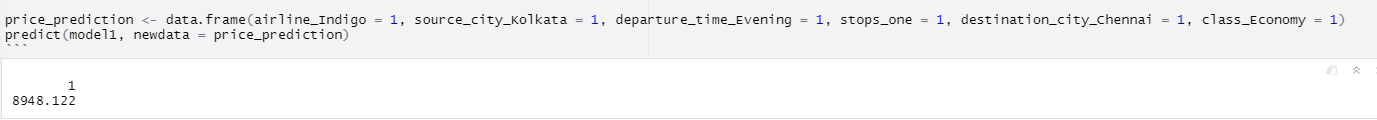
Before a prediction can be made, it must first be checked whether the model being used can return an accutrate prediction. His is done using the R-squared check. The R-squared indicates how much of the viarance in the dependent variable is explained by the xplanatory variables. The R-squared always has a value between 0 and 1, where 1 indicates the best possible model explaining all variance in the dependent variable.

**An example will be given for the assignment, all variables are taken from the dataset;**

*Stan want to fly from Kokate to Chennai and prefers to fly with the airline Indigo. He will have to fly in the evening as he is still working during the day. He also wants to save costs by flying in the conomy class and with 1 stopover.*

First of all, it is checked on the basis of the model what the R-squared is before predictions are carried out. In this case, the R-squared of the model is 0.8889, so it is good to start prediciting with it.

Then you can predict. Since a dummy column has been created for each variable here, it only needs to be checked if it is true (i.e. when a 1 is returned). It is then possible to make predictions with these variables using the predict function. The result of the assignment now shows that Stan must put down a total of 8938.12 for the flight. However, it is unkown to me in which currency this is. Probably in rupee as it is data from India. This can be seen in the image below.



It is therefore currently possible to make predictions for a ticket price bases on variables. However, you can currently only choose from an x number of variables that are included in the dataset.

**Modeling**

Last adjustment before the actual modeling was scaling the data with function *scale().* This was needed because every variable has different scale of measuring the values. When comparing different values, the measuring scales need to be the same otherwise the model will present false results.

**Price model:**

The first model was made with the ‘price’ variable as dependent. The goal of this model was to find out which variable are affected when the value of ‘price’ has increased. The result of the model are following:

**Afbeelding met tafel

Automatisch gegenereerde beschrijvingAfbeelding met tafel

Automatisch gegenereerde beschrijving**

Variable ‘Price’ has the most impact on all variables marked with ‘\*\*\*’. That are all variables except for ‘Source\_city\_Chennai’, ‘departure\_time\_evening’ and ‘arrival\_time\_Late\_Night’. The R-squared for this model is good (0.91). This indicates how much of the variance in the dependent variable is explained by the explanatory variables. Where 1 is the value representing the best possible model.

As an example, the variables are taken to explain what this means in relation to the price. This means when the value of duration is increased by 1 hour, price increases by 0.013 rupees. This is not much, there is a explanation for this: the relationship is not strong enough. The dataset shows that a 2-hour flight is more expensive than a 12 hour flight. Also when days\_left increased by 1, the price of a ticket wil decrease by 0.08 rupees.

The dummy variables are interpreted bases on the reference category: that is the dummy variable that is not in the model. Let’s take the airlines as an example; airline\_Vistara is not included in the model. In this model this means that Air\_India compared to Vistara, is on average 0.17 rupees cheaper. It can therefore be concluded that airline\_airAsia is the cheapest airline. It’s compared to airline\_Vistara 0.18 rupees cheaper, no airline is cheaper than this.

The next dummy data I looked at is the ‘source\_city’. In this case, Mumbai is the reference category here. It has therefore been omitted from the model. In this model this means that for example that flying from source\_city\_Hyderabad compared to source\_city\_Mumbai, is on average 0.06 rupees cheaper. It can be conluded that flying from the source\_city\_Hyderabad is the cheapest.

Then we are looking at the best time to depart, so when is the flight cheapest based on departing time. Here is departure\_time\_night the reference variable. This one is omitted from the model. In this model this means that for example flying in the afternoon compared to the night, is on average 0.03 rupees cheaper. It can be conluded that flying is cheapest in the afternoon.

The next topic is the number of stops during the flight. Here is stops\_zero the reference variable. In this model this means that flying with 2 stops is 0.43 rupees more expensive than 1 stop. Actually this is weird because when you make more stops the flight is actually cheaper. However, the model does not check the same flight, but also includes short flights (not the same duration), which are often cheaper. Flights with multiple stops are more expensive flights. That is why the model indicates stops with 0 stops are the cheapest.

The next is the arrival time, when is the best time to land for the cheapest possible flight ticket. Here is arrival\_time\_night the reference variable. This one is ommited from the model. In this model this means that arriving in; arrival\_time\_Early\_morning compared to the night, is on average 0.08 rupees cheaper. It can be concluded that arriving in the early morning is the cheapest. The rest of the variables are all more expensive or less cheap than in the early morning.

The next is about the place of destination. In this case destination\_city\_Delhi is the reference variable. This one is commited from the model. In this model this means that flying to Hyderabad compared to flying to Delhi is the cheapest, it is on average 0.007 rupees cheaper. All the other cities are more expensive to fly into. This also corresponds to the source city. It is also the cheapest to fly from Hyderabad.

The last one is about the chair classes so businnes flight or a economic flight. The reference variable is in this case the economic one. In this model this means that flying with businessclass compared to economic class is more expensive. On average 1.98 rupees more expensive. It can therefore be concluded that flying with economic class is a lot cheaper than flying buiness class.

To summarize everything, the cheapest variables are listed below:

**Airline:** airline\_airAsia

**Source:** source\_city\_Hyderabad

**Departure time:** departure\_time\_Afternoon

**Stops:** 0 stops (more explanation in the above section)

**Arrival time:** arrival\_time\_Early\_morning

**Destination city:** Destination\_city\_Hyderabad

**Class:** Economy

In het code below, the model and the table are created using the APA rules. Learned from the minor data driven desiscion making in business.

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

The models below are not further explained. These are also less important for the research. For the research it was important to know which variables are important to book a cheap flight. This has been achieved in the above model. Furthermore there s a very low R-squared that means the model that I tested explains very little variance. Only a very small part of all the spread in the dependent variable can be explained by the independent variable.

An example for this; when the price goes up with 1 rupee. The duration of the flight will go up with 0.08 hour. That is nice to know, but this has no addes value within the scope of the project.

**Duration model:**

**Afbeelding met tafel

Automatisch gegenereerde beschrijvingAfbeelding met tafel

Automatisch gegenereerde beschrijving**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

**Afbeelding met tafel

Automatisch gegenereerde beschrijvingDistance model:**

**Afbeelding met tafel

Automatisch gegenereerde beschrijving**

*Afbeelding met tekst

Automatisch gegenereerde beschrijving*

# 6. Data science Techniques and Tools

This chapter looks back at certain project choices. The choices of a number of tools and techniques will be discussed.

**Linear regression model:**

Within this project it was decided to use a linear regression model. This was a personal preference. This is because I gained the most experience with linear regression during the minor. This also made it more pleasant and easier for me to work with this instead of KNN or naive bayes. These methods were applied in one course within the minor, but after that I never worked with them again. So it would take a lot of time to go into this in more detail and find the right dataset for it.

**Language r:**

The same applies to the language R as to the regression model. Within the minor, I were only introduced to the language R. That is ttherefore the reason why I chose this language. To switch to Python, for example, would be quite a challange. There is no theory of this language available within our minor. This would take a lot of extra time. I therefore chose the easy way and to apply the learned language. In this case R.

# 7. Methodology

In this chapter it will be explained how I proceeded to finally arrive at the product within the chapter.

Deskresearch was used for the first part. The chapter ‘literature study’ has been written on the basis of deskresearch. The problem of which variables make airline tickets expensive has been researched before. This ensures that it becomes a deskresearch.

# 8. Conclusion

To answer the main research question: ‘What ensures that you can book a plane ticket as cheaply as possible?’. First it will have to be determined which variables are responsible for this. This in comparison with the supplied data form the dataset. This, of course, imposes a limitation on this study. Also because there are only a few unique ID’s per column (think of 6 city of sources) in the dataset.

A number of important variables have been included that can be found on the basis of deskresearch and can also be found within the dataset, namely; Dinstance, competition, flight travel type and flight timing. Unfortunately, the other variables in the literature study are not taken into account. This is partly because this a snapshot of the dataset. This means that this data is from one day (11 February) so it is not possible to do calculations using peak season. This was all done within the business understanding and dataundersting part.

Secondly, the data will have to be prepared for modelling. There is looked at; creating dummies, missing values, multicollinearties and outliers. This in order to be able to model properly and to draw the right conlusions. The dataset was already clean and this was actually the last extra check to find out the last disagreements.

In the modeling part it can be concluded that the variable ‘Price’ hast the most impact on all variables with ‘\*\*\*’ (if you look to the price model). That are all variables except for ‘Source\_city\_Chennai’, ‘departure\_time\_evening’ and ‘arrival\_time\_Late\_Night’. Subsequently, the values within these columns were examined in more detail. So I looked at which variables ensure that you can book a flight ticket as cheaply as possible. These are the variables:

**Airline:** airline\_airAsia

**Source:** source\_city\_Hyderabad

**Departure time:** departure\_time\_Afternoon

**Stops:** 0 stops

**Arrival time:** arrival\_time\_Early\_morning

**Destination city:** Destination\_city\_Hyderabad

**Class:** Economy

A prediction model has also been made. Where values can be given. A prediction of the price of the flight ticket will then be rolled out.

# 9. Reflection

Looking back at my goal of the project report: map out my learning process & apply techniques from the learned theory. I can conclude that this has been successful for this project. The entire cycle was run on the basis of the CRISP model. The theory learned in the minor has been applied to this flight prediction dataset. I am also happy with the final results; a linear regression model and a prediction model.

As a bim-mer (name for bachelor study, student) I also made steps in the project. By getting started with datascience techniques and tools I got a somewhat broader view instead of only focussing on the business side of IT. This naturally benefits my personal development within ICT. At the moment I am not sure whether I want to do anything else with data science in the form of a master degree. I see this as something potential after my bachelor.

In terms of the project, the modeling was better than expected in my opinion. Despite some help with reading the model, it was manageable. I expected that I would run into a little more problems since I was now given a non-pre-assigned dataset but had to find one myself. I also found it difficult at first to really make a start and where to start. I did notice that CRISP had a positive influence on this. By literally writing out these phases within the report, I created something to hold on to for myself, so that I could see acactly what I had to do for each chapter. Overall, I consider this project to be a success.

# 10. Bibliographie

BATHWA, S. (2022, 12 21). *Kaggle.com*. From Kaggle.com: https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

Intermiles. (2020, 10 22). *Top Factors That Determine Flight Ticket Prices*. From Intermiles: https://www.intermiles.com/blog/top-factors-that-determine-flight-ticket-prices