# Documentation

## 1. Motivation

### *1.1 For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?*

According to research of Eurostat (2020) 1,034 million Europeans travelled by plane in 2019. This is an increase of 3.8% compared with 2018. Corona has shut down the aircraft industry for a while, but the industry is currently getting back on track. Because this market has so many customers, there are also many suppliers. It is estimated that there are around 5.000-5.500 airlines across the world, of which there are 700-800 commercial airlines operating scheduled flights globally. According to research airlines have to focus on customer satisfaction and customer experience to differentiate itself from other competitors in this market (Ban & Kim, 2019). Studies have shown that analysing online review data has several advantages for the customers' satisfaction and the way the customers experience the airlines. These advantages are:

- It is an inexpensive way to gather information as reviews are often already available on websites or apps (Liau & Tan, 2014).
- Reviews from customers are considered as trustworthy (Brochado et. Al, 2019).
- Reviews are fast, which means that reviews are often put online within days of a customer's "purchase" (Brochado et. Al, 2019).

These advantages show that gaining data from reviews of customers can be an interesting way to analyse and improve customer satisfaction. However, there is no dataset available where airlines can analyse all reviews of their airline. Analysing reviews from website is extremely time consuming because little to no filters can be applied and therefore it is very difficult to make segments. Scraping the reviews from a website to make a dataset, can therefore be very useful for airlines.

There are plenty of websites on the internet where reviews of airlines are collected. Several websites have been analysed and the website airlinequality.com has been chosen as the best option for scraping because of the following reasons: * Airlinequality.com gives customers the opportunity to verify their flight though by uploading their boading pass or ticket. Airlines can therefore be sure that the verified reviews are written by people who have actually used the airline. As a result, the dataset contains more credible data. * Airlinequality.com lets customers rate the airline on several different variables which already have been mentioned above. This gives the airline star ratings on various variables which gives the airline insight into its good and bad points. * Airlinequality.com is an international website, which means that it can get reviews from customers of all countries.

The websites that have been analysed besides airlinequality.com are mentioned below and it is given why these website will not be used to scrape reviews. * Tripadvisor.com. This website lets the customer rates on different variables but does not give verified reviews, which means that the website does not check whether the customer has actually been on the flight he is reviewing. This allows people to write fake reviews, this reduces the reliability of the dataset. * Flight.report.com. This website lets customer also rate on different variables but each review is published on a separate page. When scraping these reviews, each page should be scraped individually to get all the reviews. This is extremely time consuming and therefore these website has not been chosen to scrape.

### *1.2 Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

The dataset that is available after scraping the website airlinequality.com has been scraped by a projectgroup of the course Online Data Collection and Management. This course has been given by Hannes Datta at Tilburg University and is part of the master Marketing Analytics.

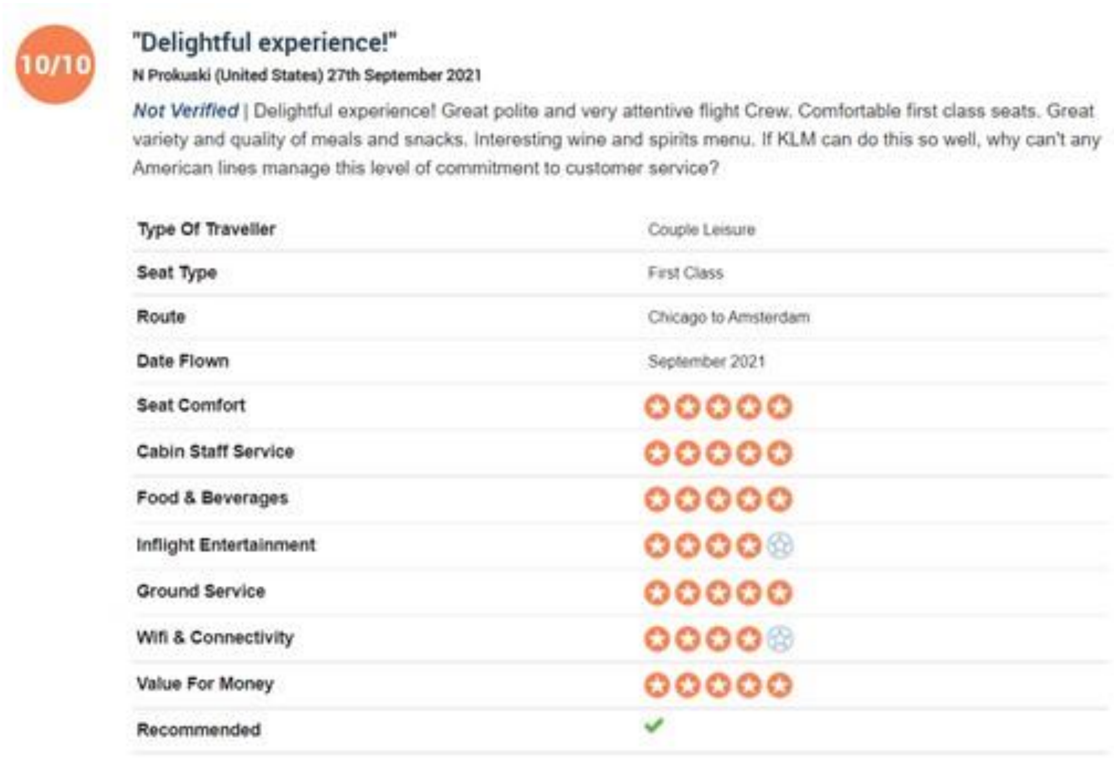
### *1.3 Who funded the creation of the dataset?*

No external partner directly funded the creation of the dataset. But the possibility of compiling this dataset comes from the teaching materials of Tilburg University. They provided teaching materials that ensured that project group members had enough information to scrape the airlinequality.com website.

## 2. Composition

### *2.1 What do the instances that comprise the dataset represent?*

Each instance of the dataset represents a review of a traveller of KLM. The instances gives information about the review but also about the traveller who wrote the review. In figure 1.2.1 an example of a review on airlinequality.com is given. In figure 2.1.1 an example of a review on airlinequality.com is given.



10/10

"Delightful experience!"

N Prokuski (United States) 27th September 2021

*Not Verified* | Delightful experience! Great polite and very attentive flight Crew. Comfortable first class seats. Great variety and quality of meals and snacks. Interesting wine and spirits menu. If KLM can do this so well, why can't any American lines manage this level of commitment to customer service?

| | |
|---|---|
| Type Of Traveller | Couple Leisure |
| Seat Type | First Class |
| Route | Chicago to Amsterdam |
| Date Flown | September 2021 |
| Seat Comfort | ★★★★★ |
| Cabin Staff Service | ★★★★★ |
| Food & Beverages | ★★★★★ |
| Inflight Entertainment | ★★★★☆ |
| Ground Service | ★★★★★ |
| Wifi & Connectivity | ★★★★☆ |
| Value For Money | ★★★★★ |
| Recommended | ✓ |

figure 2.1.1: example of a review

## 2.2 How many instances are there in total?

In total the dataset consist of 1.230 instances, in other words, 1,230 reviews have been processed in the dataset. This does not mean that all those reviews were written by 1.230 different passengers. According to the dataset 1,102 unique writers wrote the reviews. It must been said that, before handing in the review, the traveller can choose that a screen name is visible instead of their own name. As a result, it cannot be guaranteed that the dataset actually contains 1,102 unique writers since writer could write different reviews but all with another screen name.

| | |
|---|---|
| Reviews: | 1230 |
| Unique writers: | 1102 |

## 2.3 Does the dataset contain all possible instances or is it a sample of instances from a larger set?

The dataset consist of all the reviews that the webite airlinequality.com has of KLM. So no sample was made of the reviews of the website. But not all KLM travelers write a review on airlinequality.com after they flew with KLM. So in that case it can be said that the dataset is a sample of a larger dataset. The dataset contains of the reviews that have been published between 21-03-2013 and 08-10-2021. In the table below it is given how much reviews are available in the dataset per year and how much passengers have travelled with KLM in that year.

| *Year* | *Reviews of flight in dataset* | *Total passengers of KLM* |
|---|---|---|
| Before 2015 | 344 | - |
| 2015 | 126 | - |
| 2016 | 155 | 30,399,000 |
| 2017 | 150 | 32,689,000 |
| 2018 | 154 | 34,170,000 |
| 2019 | 188 | 25,092,000 |
| 2020 | 70 | 11,231,000 |
| 2021 | 43 | ? |
| *Total* | 1230 | - |

(Mazareanu, 2021)

In the recent years, only a very small proportion of KLM travelers have left a review at airlinequality(e.g. in 2019 0.00075%). So a sample of the entire population was taken, but it was not selected randomly. The members of the population who left a review on airlinequality.com were selected for the sample. When research is done with the dataset, they represent the population. In the following tables it is shown what the dataset consists of and whether the dataset is representive.

In the following table it has been shown whether the dataset is representative geographically for all travelers of KLM.

| *Country* | *frequency in dataset* | *% of the dataset* |
|---|---|---|
| United Kingdom | 271 | 22.0% |
| Netherlands | 192 | 15.6% |
| United States | 129 | 10.5% |
| Canada | 87 | 7.1% |
| Germany | 64 | 5.2% |
| Australia | 36 | 2.9% |
| Switzerland | 30 | 2.4% |
| Singapore | 26 | 2.1% |
| Other countries | 395 | 32,1% |

In the dataset there are 78 different countries mentioned. In 2019, KLM flew to 112 different countries (AirfranceKLM, 2021). Assuming that at least one person in these countries has flown with KLM, there are still many countries that are not represented in the dataset. Mainly travellers from the United Kingdom, Netherland and the United States are represented in the dataset. Almost half of the people in the dataset are from these three countries.

| *Type of Traveller* | *Freq. in dataset* | *% of the dataset* |
|---|---|---|
| Solo Leisure | 353 | 28.7% |
| Couple Leisure | 225 | 18.3% |
| Business | 178 | 14.5% |
| Family Leisure | 131 | 10.7% |
| Missings | 343 | 27.9% |
| *Total* | 1,230 | |

The dataset mainly represents the Solo Leisure travellers and the Couple Leisure travellers. But the business travelers and the family travelers are also represented in the dataset.

| *Seat Type* | *Freq. in dataset* | *% of the dataset* |
|---|---|---|
| Economy Class | 895 | 72.8% |
| Business Class | 244 | 19.8% |
| Premium Economy | 88 | 7.2% |
| First Class | 3 | 0.02% |
| *Total* | 1,230 | |

The dataset contains of namely reviewers who travelled in Economy Class and Business Class. The dataset has almost no reviews of travellers Premium Economy and First Class. When information is needed for Premium Economy and First Class travelers, this dataset is not suitable.

## 2.4 What data does each instance consist of?

The dataset consists of raw data. Meaning that the data has been unproccesed. All the data has been scraped from the qualityairline.com and put into the dataset without without any data being removed or added to the data. The information that has been given in the dataset has been published in the same way on the website of airlinequality.com. Per review/reviewer the following information has been given:

- Column 1. In the first column the overall rating of the reviewer on KLM has been given. This rating can contain a number between 1 and 10, with 1 being the lowest score and 10 being the highest score. There are no missing in this column, since this was a required part to complete the review.
- Column 2. In column 2 the name of the reviewer has been given. It must be said that every reviewer is given the opportunity to indicate that he or she does not want her

own name on the review. There are no missing in this column since airlinequality.com changes the names of the people who do not want their name appeared on the review, to a fake name.

- Column 3. Column 3 gives the title of the review. This title of the review is a part of the written review of the reviewer, which describes an important experience of sums up the whole review. There are no missings in this collums

- Column 4. Column 4 gives the date of the day the reviews were published in the form of YYYY-MM-DD. The oldest review of the data set is from 08-06-2021. This means that all the data in the dataset has been published after that date.
- Column 5. In column 5 it is given if a reviewer has verified his trip or not. People who have verified their trip (by uploading their boarding pass/ticket) are given "Trip Verified". Those who are not are labelled with "Not Verified".
- Column 6. Column 6 gives the country in which the reviewer lives. There are no missings in this column since this question was a required part to complete the review.
- Column 7. In column 7 it is given in which aircraft the review flew, if they knew or wanted to answer the question. The reviews who did not answered that question are labelled as missing.
- Column 8. Column 8 gives the flight the reviews were on in this form "take-off location" to "landing location". In this column there are also no missings since this was a required part of the review.
- Column 9. In column 9 it is given which type of traveler the reviewer was on the flight he reviews. The 4 different answers that could be given were: Solo leisure, Business, Couple Leisure and Family Leisure. In this column there are also no missings since this was a required part of the review.
- Column 10. Column 10 gives the seat type of the reviewer. The four possible answers the reviewer could give were: Business Class, Economy Class, First Class and Premium Economy. In this column there are also no missings since this was a required part of the review.
- Column 11. In column 11 the month and year has been given of the flight of the reviewer. These answers are given in the form month-year.
- Column 12. In column 12 the answer of the rating for "Value for Money" has been given. The rating can contain of a number between 1 and 5, with 1 being the lowest score and 5 being the highest score. In this column there are also no missings since this was a required part of the review.
- Column 13-18. In the columns 13-18 the ratings are given for respectively "Seat Comfort", "Cabin Staff", "Food & Beverages", "Inflight entertainment", "Wifi & Connectivity" and "Ground Service". The rating can contain of a number between 1 and 5, with 1 being the lowest score and 5 being the highest score. Since these ratings were not a mandatory item of the review, there are missings for each of these

columns. The exact numbers of missing are given in the paragraph "Data inspection per entity".

- Column 19. In the last column it is given whether the review would recommend the airline, in this case KLM, or not. In case the reviewer would recommend the airline, a 'yes' has been given. In case the reviewer would not recommend the airline, a 'no' has been given.

### 2.5 Is there a label or target associated with each instance?

Not applicable.

### 2.6 Is any information missing from individual instances?

There are several instances that contain missings in the dataset. There are several reasons for this:

1) It appears that until April 2015 the traveller's route, traveler type and date the reviewer flew were not asked when a reviewer wrote a review. No data is available untill this date. After April 2015 no missings were found in these variables.

2) Reviewers could fill in the type of aircraft if they knew the type of aircraft. When people did not fill in the type of aircraft, is has been registered as a missing.

3) For the rating on several points only the overall rating and the rating on "Value for Money" were mandatory. For the other ratings, reviewers could choose whether they rated the variable or not. When people did not review the variable, the variable is shown as a missing.

For the variables that contain missings, the number of missings are given below:

- Type of aircraft → 715 missings
- Rating on Seat Comfort → 36 missings
- Rating on Cabin Staff Service → 36 missings
- Rating on Food & Beverages → 118 missings
- Rating on Inflight Entertainment → 366 missings
- Rating of Wifi & Connectivity → 1029 missings
- Rating of Ground Service → 368 missssings
- Route the plain flew → 343 missings * Type of traveller → 343 missings * Date of the flight → 344 missings

One traveler, has entered the route and the type of traveler, but not the day that he flew. How this is possible, is not clear.

### *2.7 Are relationships between individual instances made explicit?*

Not applicable.

### *2.8 Are there recommended data splits (e.g. training,development/validation, testing)?*

For the data set that has now been given, splitting is not yet recommended, since in the data set some subgroups only have little representation and when the data set is split, this group will only become smaller. The code that has been given for scraping the data can be used again to scrape new reviews. if, for example, new data is added every year from new reviews, it is recommended to split the dataset into a testing and a training dataset. In that way new models for analysing data could be tested easiy with the testing data and then could be applied for the training data.

### *2.9 Are there any errors, sources of noise, or redundancies in the dataset?*

There are no errors, sources of noise or redundancies in the dataset.

### *2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)?*

The dataset is rely on the external resource www.airlinequality.com and especially on the KLM reviews page Since the data in the dataset is scraped from airlinequality.com, an extension of the dataset depends on whether: 1) The customers keep handing in reviews on this website and 2) Whether this website will continue to keep the reviews available online. But because it is expected that the reviews are largely intended for other travelers, the second point will certainly not endanger the development of the dataset. Since the whole dataset is derived from the external source, there is no dataset without the external source. This external source does not charge any fees or licenses to scrape the data on the website.

The latest version of the dataset (which contains data up untill 14-10-2021) can be found in the data folder.

### *2.11 Does the dataset contain data that might be considered confidential?*

No, the data available is also publicly available on the internet so no confidential data is included in the dataset.

### *2.12 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?*

The title of the review might be considered as offensive. People can write positive reviews as well as negative reviews. It is therefore possible that, in this case KLM, is offended by the title of the review.

### *2.13 Does the dataset relate to people?*

Yes, the data in the dataset relates to people since they have wrote the reviews and some information is given about these people.

### 2.14 Does the dataset identify any subpopulations?

No, subpopulations have been made in the dataset. But segmentation can be done in several ways to obtain subpopulations from the dataset. Depending on the researcher's goal, subpopulations can be created based on:

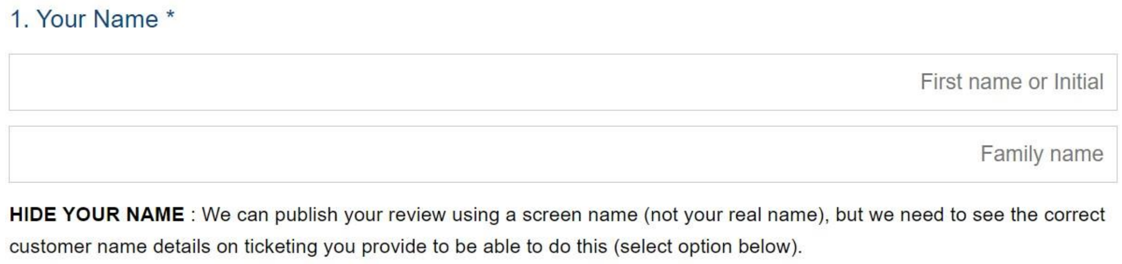
- The country they live in
- The route the flew
- The type of traveller they were
- The type of seat

In paragraph 2.3 it is given how many of each of these subgroup are available in the dataset.

### 2.15 Is it possible to identify individuals, either directly or indirectly from the dataset?

The following things are known about each review: * Name, in case the reviewer gave premission. If the reviewer did not give permission to publish the name, a fake name appeared. It can not be seen whether the name of a review is their real name or their fake name. * Country the live in. * Flight the reviewer took but not the exact date. * Seat the reviewer took. * The type of traveller the reviewer was.

With the information above reviewers could be identified when they used their own name. Reviewers had the choice whether they wanted their own name published or a screen name. this. In image 2.15.1 you can see what the reviewers saw when they filled in the review.



1. Your Name *

First name or Initial

Family name

**HIDE YOUR NAME** : We can publish your review using a screen name (not your real name), but we need to see the correct customer name details on ticketing you provide to be able to do this (select option below).

image 2.15.1: name of the reviewer

Identifing the reviewers can be done by the airline for which the review was written. They can check in their own database whether the name of the reviewer has actually taken the flight that he says he has taken in the review. If, according to the database, the reviewer did not take the flight indicated in the dataset, it can be concluded that the traveler did not use their own name. If the real name is not used, it is not possible to trace the traveler. The exact date of the flight is not given, so the airline cannot see which flight the traveler has taken.

### *2.16 Does the dataset contain data that might be considered sensitive in any way?*

No, no data is included that may be considered sensitive in any way.

## 3. Collection Process

### *3.1 How was the data associated with each instance acquired?*

The data that is collected is reported by subjects. Subjects leave a review about a specific airline, in this case KLM, on www.airlinequality.com. The verification of the data happens via the verification of the flight. Subjects need to fill in their flight details in order to verify that they actually flew with the airline that they are reviewing. If their flight is verificated, the review will be verified. In our dataset our also flights that are not verified. However, these reviews can be filtered out when necessary while analyzing the dataset.

### *3.2 What mechanisms or procedures were used to collect the data*

For the collection of this data, BeautifulSoup is used instead of Selenium because there is no need to scroll through the pages. Furthermore, the Java elements that are present on the website are not necessary for scraping so therefore BeautifulSoup is used.

In order to scrape www.airlinequality.com, there are a few steps that a potential researcher needs to do manually. First, select the airline that is of interest. Airinequality provides a list of airlines from A to Z where you can select the airline or airlines that you are interested in. Second, select the period that should be investigated. Based on this period, the researcher needs to determine the number of pages that covers this period and this number of pages should be implemented in the code.

### *3.3 If the dataset is a sample from a larger set, what was the sampling strategy?*

As far as www.airlinequality.com is concerned, we scraped every possible review for KLM for our dataset. Therefore, we did not use a sampling strategy.

### *3.4 Who was involved in the data collection process and how were they compensated?*

The data is entirely collected by our research group from Tilburg University. Therefore, there is no need for compensation.

### *3.5 Over what timeframe was the data collected?*

The dataset contains observations from the 21st of March 2013 until the 12th of October 2021 20:57.

### *3.6 Were any ethical review processes conducted?*

No ethical review processes were conducted.

### *3.7 Does the dataset relate to people?*

The dataset relates to people, since the reviews of different people are collected.

### *3.8 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?*

The collected data is obtained from www.airlinequality.com which can be best described as third party. This website collected the data from the reviewers.

### *3.9 Were the individuals in question notified about the data collection?*

Individuals were not notified about the data collection.

### *3.10 Did the individuals in question consent to the collection and use of their data?*

Individuals did not explicitely consent to the collection of the data. However, they allowed their review to be shown on www.airlinequality.com which makes it visible to anyone that it interested in this kind of data.

### *3.11 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?*

Not applicable.

### *3.12 Has an analysis of the potential impact of the dataset and its use on data subjects?*

Data subjects are not negatively affected by analysis of this dataset. Instead, it might even be beneficial for data subjects since airlines can determine with this data which improvements they need to make in order for all customers to be satisfied.

### *3.13 Any other comments?*

Not applicable.

## 4. Preprocessing/cleaning/labeling

### *4.1 Was any preprocessing/cleaning/labeling of the data done?*

No preprocessing was done before and after scraping the data. It has been decided that the names of the reviewers are visible in the dataset since the reviewers are visible with their own name on the website airlinequality.com and had the choice to make the review anonymous if they wanted to.

No instances were deleted when they contained one or more missings. This is because they did enter information from other variables that may also be of interest to users of the dataset.

## 5. Users

### *5.1 Has the dataset been used for any tasks already?*

The dataset has not yet been used by other people or researchers besides the projectteam that has scraped the information of the dataset.

### *5.2 Is there a repository that links to any or all papers or systems that use the dataset?*

Since the dataset has not yet been used, there are no papers or systems that use the dataset.

### *5.3 What (other) tasks could the dataset be used for?*

The dataset obtained after scraping the website airlinequality.com is mainly made for KLM. As already described in the motivation, it is very valuable for airlines to observe reviews to enhance the customer experience of a customer. This dataset is valuable because it saves valuable time because all information is contained in one data file. This allows marketers to easily create segments and find out what the wishes and needs of these segments are.

### *5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No

### *5.5 Are there tasks for which the dataset should not be used?*

The dataset should not be used to contact the travellers in the dataset, in case they used their own name on the review, for commercial or other reasons.

## 1.3 Collection process

### *Technical extraction plan*

The website that is scraped is, as stated before, www.airlinequality.com. This is a website and therefore we could either use BeautifulSoup or selenium in order to scrape data from the website. We decide to use BeautifulSoup for scraping. The reason that we decided to do this is mainly because we do not need the Java elements of www.airlinequality.com in order to properly scrape the website. Also, since www.airlinequality.com show all the reviews instantly without the need for scrolling throughout the page. Furthermore, all the elements that we want to scrape from www.airlinequality.com are HTML-elements. Therefore, BeautifulSoup works perfectly for scraping www.airlinequality.com. There are multiple entities that are scraped from www.airlinequality.com. First, in order to scrape all the reviews for one airline during COVID-19, we initially wanted to determine which pages of reviews could represent the COVID-19 period. As it was hard to determine these pages with python, we decided to scrape all the reviews of KLM. With R this data set can than be filered to cover the COVID-19 period. We included all the reviews up to 2021-14-10 11:38. For KLM this means we scraped 123 pages of reviews. The number of pages depends on the airline, but can be based on the total pages function provided in the jupyter notebook. Second, we also determined to scrape multiple airlines. The airlines can be manually chosen from the airline list saved in the jupyter notebook.

In order to scrape www.airlinequality.com, there are a few steps that a potential researcher needs to do manually. First, select the airline that is of interest. Airine quality provides a list of airlines from A to Z where you can select the airline or airlines that you are interested in. The name of this airline should than be added to the airline list in the jupyter notebook. Second, select the period that should be investigated. Based on this period, the researcher needs to determine the number of pages that covers this period and this number of pages should be implemented in the code. Alternative, the researched can use the total pages function in the jupyter notebook to generate the total pages of reviews for a specific airline.

The data set is finalized on 2021-14-10 11:38. Therefore, no reviews that are submitted after this data are visible in the dataset.

### *Legal and ethical concerns*

There are no legal and ethical concerns when scraping this data. The people that are submitting their reviews on this website know that their information is visible. Furthermore, the collected data is, besides the name of which they can indicate whether they want to use their own name, fully anonymous.

## About

This research is carried out in implementation of the course Online Data Collection and Management. This is a part of the Master program of Marketing Analytics

## Sources

Ban, H.-J., & Kim, H.-S. (2019), Understanding Customer Experience and Satisfaction through Airline Passengers' Online Review. Sustainability, 11(15), 4066. doi:10.3390/su11154066

Brochado, A., Rita, P., Oliveira, C. and Oliveira, F. (2019), "Airline passengers' perceptions of service quality: themes in online reviews", International Journal of Contemporary Hospitality Management, Vol. 31 No. 2, pp. 855-873. https://doi.org/10.1108/IJCHM-092017-0572

Eurostat. (2020 November), Air transport statistics. accessed at October 2021, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Air_transport_statistics#Progressive_growth_in_air_transport_of_passengers_in_the_course_of_2019

Mazareanu , E. (2021), Annual number of passengers carried by KLM Royal Dutch Airlines from 2014 to 2020, accessed at Octboer 2021, from https://www.statista.com/statistics/734736/annual-number-of-passengers-carried-byklm-royal-dutch-airlines/

Yee Liau, B. and Pei Tan, P. (2014), "Gaining customer knowledge in low cost airlines through text mining", Industrial Management & Data Systems, Vol. 114 No. 9, pp. 13441359. https://doi.org/10.1108/IMDS-07-2014-0225