Flight Cancellation Data  
*Sabine Abrahamse, Federico Fiorani, Kevin ‘t Jong, Lorenzo Taddei, Marton Tomka*

Derive

# 1. Motivation[[1]](#footnote-1)

***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? Please provide a description.*

Because of the pandemic, people have significatively changed their flying habits. Flights are being purchased in the proximity of the departure date as is not easy to make long term plans anymore. From the airline perspective, this led to increasing difficulties in managing the flight traffic. This has resulted in a rise in the flight cancellation rate in all the international airports and for all the best-renowned airline companies. The data regarding this problem are not accessible to users or, if published, are only referring to the past 2 days. Our mission is to collect all the details regarding cancelled and delayed flights in the biggest international airport every day until the creation of a large and very precious dataset. We think that through the observation of historical data regarding this phenomenon, we will be able to determine the risk of cancellation for upcoming flights.

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

FlightAware is a digital aviation company and operates the world's largest flight tracking and data platform. With global connectivity to every segment of aviation, FlightAware provides over 10,000 aircraft operators and service providers as well as over 13,000,000 passengers with global flight tracking solutions, predictive technology, analytics, and decision-making tools.

***1.3*** *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

FlightAware receives data from air traffic control systems in over 45 countries, FlightAware's network of ADS-B ground stations in 195 countries, Aireon global space-based ADS-B, and datalink (satellite/VHF) via every major provider, including ARINC, SITA, Satcom Direct, Garmin, and Honeywell GoDirect. There is no extra funding.

# 2. Composition

***2.1*** *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of in- stances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The instances that comprise the dataset represent cancelled flights. There are multiple types, namely the airline in charge of the cancelled flight and the origin airport of the cancelled flight.

***2.2*** *How many instances are there in total (of each type, if appropriate)?*

The number of instances are the same for both the airline and the origin airport. The number of instances changes each day and it depends on the number of flight cancellations. There are also unknown airlines in the data, but the scraper does not take them into consideration. The approximately 200 airlines that are listed in the code are analyzed.

***2.3*** *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset contains all possible instances. With between 1500 and 2000 cancelled flights each day, the scraper can run them in about 30 minutes. This means the scraper does not overload the computer. A sampling run has also been done, to see if the code was running smoothly, where only one page per airline was collected. The code for this can be found in the script.

***2.4*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.*

Each instance consists of text. The text provides the name of the airline and the name of the origin airport, which makes it possible to analyze also geographically.

***2.5*** *Is there a label or target associated with each instance? If so, please provide a description.*

The label that is associate with each instance are the 200 airlines that are analyzed. The target is the origin airport that is specific for each flight.

***2.6*** *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

There are a few unknown airlines with cancelled flights each day, which means this data is missing for the instances. Also, for these unknown airlines, there is no association with an origin airport, which makes it hard to use this data. They usually account for only 1 cancelled flight per airline each day, which is less than 2%. These airlines are not categorized, so the focus is on the remaining 200 known airlines.

***2.7*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

The relationship between the name of the airline and the origin airport is made explicit in each case. The origin instances are organizes into the airline instances, like a subcategory. This is also the way the website is structured, so the extraction process makes these explicit relationships easy to maintain.

***2.8*** *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

There are no recommended data splits. All the collected data is pulled together in a single csv-file.

***2.9*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The dataset relies on the number of flights cancelled all over the world and the website FlightAware.com collects the information on these flights. Data on flights cancelled everyday will be available, but there are no guarantees that FlightAware.com runs 24 hours a day without crashing. The code scrapes data from the previous day, so unless the website stops working for more than 24 hours, the data can be collected. The structure of the website could change and this requires an update of the code. The data of the website is historically available up to one day before. In this way, a historical data set is created based on everyday scraping, so predictions can be made. There are no restrictions to get access to the data, such as purchase licencees to scrape the website. The data is easily accessible using the "selenium" package. There is no guarantee this will not change in the future.

The link to the website of cancelled flights: <https://uk.flightaware.com/live/cancelled>

The link to the website of cancelled flights, one day before: <https://uk.flightaware.com/live/cancelled/yesterday>

An example of the website that shows the cancelled flights of a certain airline, in this case China Airlines: <https://uk.flightaware.com/live/fleet/CCA/cancelled>

***2.10*** *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

The data on cancelled flights is public information. The data is easily available on the internet. FlightAware.com only collects and summarizes this data.

***2.11*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

The dataset might cause anxiety for passengers who travel a lot by air. The goal of the website, however, is to provide information for people who need data on cancelled flights to for example estimate the risk of cancellation for a flight they are planning.

***2.12*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The data contains solely about flights.

# 3. Collection Process

***3.1*** *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The instances were collected beforehand by creating a list of all the airlines displayed on the website. The name of the airlines was used to spot their links to all the cancelled flights for each airline. A "for loop" was used over all the airlines. All the pages with flights were accessed by clicking on "Next 20" on each specific page. The airlines' names are directly observable on the main page, whereas specific flight data is scrapable from specific links for each airline. All the data is scraped in textual form from FlightAware.com. The calculations that can be done are based on the datasets of the scraped data, validated by the source FlightAware.com.

***3.2*** *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

The data is collected with webscraping via python, which means that the data is directly scraped from the website, making use of html scripts and tags. The procedures have been validated by checking the correspondence of the data in the scraped dataset with the data present on the website in html form.

***3.3*** *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

Initially, a deterministic sampling was used. This means data was collected only from the first page of each airline to understand whether the code was correctly retreiving data from the website.

***3.4*** *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

The team working on the webscraping project is composed by five students following the Online Data Collection and Management (oDCM) course at Tilburg University, with the support of the professor of this course. The work is aimed at practicing aqcuired knowledge, thus no financial compensation was used.

***3.5*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

The data has been collected since the 13th of October. With the use of Task Scheduler, the data will be collected daily for an indefinite period of time.

***3.6*** *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No review processes were done by external agents.

# 4. Preprocessing, cleaning, labeling

***4.1*** *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

The program R was used to clean the dataset, where missing values were deleted.

***4.2*** *Was the “raw” data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

Link to raw data on the github: …

***4.3*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

The software that was used to clean the data is R, which is publicly available. R can be downloaded through this link:

<https://cran.r-project.org/>

# 5. Uses

***5.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

The dataset has been used to make analyses in R.

🡪 …

***5.2*** *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

…

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

Some examples of tasks that can be done with the dataset are:

* Analyze which airlines are most likely to have their flights cancelled
* Analyze which origin airports are most likely to cancel flights
* Predictions for passengers about the likelihood of the cancellation of their future flight

***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? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

There are approximately 200 airlines analyzed. There is a possibility that the dataset that was created is not 100% complete since there are more than 200 airlines in the world. However, the largest ones of the world are all included.

***5.5*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset should not be used as a sole method of choosing between potential airlines for a passenger while purchasing flight tickets. It should only be one factor, and the passenger should consider other factors as well, like ticket prices, service on the flight, flight time, etc.

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)