**Motivation for Dataset Creation**

**Why was the dataset created?**

This dataset was collected by the Bureau of Transportation Statistics (BTS) to specifically track the on-time performance, describing the arrival date for non-stop domestic flights by major air carriers for the year of 2008,which captured the merging and acquiring of certain airlines and carriers [1].

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

One task that this database could be used for would be to determine whether on-time performance varies by airline or time of the quarter, month, week – largely useful for flight booking agencies.

Another task could be to determine the effect that these mergers and acquisitions has had on the airplane industry. An article on the Dallas Business Journal mentions possible effects using similar, yet more specific data over the last decade [2]. However, I do not believe this specific dataset would be the best fit to prove such a pointed argument.

**Who funded the creation of the dataset?** The creation of this dataset was carried out and funded by the Bureau of Transportation Statistics.

**Any other comments?**

This dataset is updated monthly and annually, where its findings are posted on the Bureau of Transportat-ion Statistics for public viewing. Though, only this particular dataset for 2008 was listed on Enigma Public [3].

**Dataset Composition**

**What are the instances?**

The instances included in this dataset are: individual flight numbers and associated on-time performance information for that airline and carrier.

**Are relationships between instances made explicit in the data?**

The relationships between the instances can be assumed but are not particularly explicit as there is not much information provided as to how the dataset was put together. What is useful in the absence of this information is the variable descriptions [4]

**How many instances of each type are there?**

There are 112 fields and 2,116,761 rows (instances).

**What data does each instance consist of?**

Each instance consists of departure and arrival delay, origin and destination airports, flight numbers and actual departure information, et cetera. The instances seem to be related to subpopulations identified as airline and carrier. However, it is hard to tell what the distribution is, seeing as how large the dataset is.

**Is everything included or does the data rely on external resources?**

Though not directly mentioned, one can assume that the data does not rely on external sources, unless trying to interpret over time in years vs quarterly, monthly, weekly as described in this dataset. If any external resources, it would include the BTS website.

**Are there recommended data splits or evaluation measures?**

Due to the dataset being so large, it might be more useful to examine a set number of instances for each airline and carriers, seeing as there are over 2,000,000 instances within this dataset. One approach that might be used would be random sampling.

**What experiments were initially run on this dataset?**

A complete summary of this data was not provided. There was a single chart that displayed the datasets findings that described the national aviation system delay as a reason for delay [5] The chart was rather confusing and not very intuitive for reading and interpreting meaning. Additionally, the full Air Travel Consumer Report for 2008, in which this data was used for, is archived on the website [6].

**Data Collection Process**

**How was the data collected?**   
The airlines report the causes of delay in broad categories that were created by the Air Carrier On-Time Reporting Advisory Committee.

Unknown.

**Who was involved in the data collection process?** The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers.

**Over what time-frame was the data collected?**

BTS began collecting details on the causes of flight delays in June 2003.

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

Unknown. It is not explicitly said how the data is associated with each instance. One can only assume it is by flight.

**Does the dataset contain all possible instances?**

This dataset does not contain all possible instances, as it only records domestic flights through large air carriers, which excludes international flights and flights through small air carriers.

**If the dataset is a sample, then what is the population?**

The dataset is a sample of domestic flights of large air carriers in the United States, opposed to international flights through smaller carriers.

**Is there information missing from the dataset and why?**

There does not appear to be information missing from the dataset, as it has already been thoroughly cleaned by BTS for publishing in the monthly consumer report.

**Are there any known errors, sources of noise, or redundancies in the data?**

There does appear to be any known errors or noise, as there is strict coding of what is considered a late flight, which is a delay of less than 15 minutes.

**Data Processing**

**What preprocessing/cleaning was done?**

Categorizing the delays and identifying the cause of the delay, the time, and if its delay was considerable enough to be identified as a delay of more than 15 minutes. Considering this data was collected by BTS, the data processing/cleaning will have happened before presented and done thoroughly before brought to the public.

**Was the “raw” data saved in addition to the preprocessed/cleaned data?**

Summary statistics and raw data are made available to the public at the time the Air Travel Consumer Report is released.

**Is the preprocessing software available?**

The processing software used was TransStats, which is still available.

**Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?**

Yes. BTS used this data, along with other data related to consumer travel in their Air Traveler Consumer Reports that is published annually.

**Dataset Distribution**

**How is the dataset distributed?**

The dataset is distributed on the Enigma Public website. Though the trends are highlighted on BTS in the monthly and annual reports, the specific dataset is not provided on the site.

**When will the dataset be released/first distributed**?

The dataset is first released on the BTS website, along with the monthly and annual consumer report. The report is usually issued during the second week of each month.

**What license (if any) is it distributed under?**

Unknown.

**Are there any fees or access/export restrictions?**

Unknown.

**Dataset Maintenance**

**Who is supporting/hosting/maintaining the dataset?**

The dataset belongs to the Bureau of Transportation Statistics and is responsible for support and maintenance of the dataset.

**Will the dataset be updated? How often and by whom?**

The datasets are updated for that year, at least as it appears to have been on Enigma Public. Updates occur monthly and annually. For one particular dataset, after the annual report is released, it would not make much sense to update, if possible.

**If the dataset becomes obsolete how will this be communicated?**

The dataset is communicated by year. Within the BTS website, it appears that after a decade, such information is archived.

**Is there a repository to link to any/all papers/systems that use this dataset?**

Yes, BTS publishes all of the reports and papers to its website for public viewing. <https://www.transportation.gov/individuals/aviation-consumer-protection/air-travel-consumer-reports>

**If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?**

Unknown.

**Legal & Ethical Considerations**

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?**

The dataset does not relate to people.

**If it relates to other ethically protected subjects, have appropriate obligations been met?**

The dataset does not relate to ethically protected subjects.

**If it relates to people, were there any ethical review applications/reviews/approvals?**

The dataset does not relate to people.

**If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?**

The dataset does not relate to people.

**If it relates to people, could this dataset expose people to harm or legal action?**

The dataset does not relate to people.

**If it relates to people, does it unfairly advantage or disadvantage a particular social group?**

The dataset does not relate to people.

**If it relates to people, were they provided with privacy guarantees?**

The dataset does not relate to people.

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)?**

The dataset does not relate to people.

**Does the dataset contain information that might be considered sensitive or confidential?**

No.

**Does the dataset contain information that might be considered inappropriate or offensive?**

No.

References

[1] Bureau of Transportation Statistics. (n.d.) Data Profile. United States Department of Transportation. Retrieved from <https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&DB_URL>=

[2] Hethcock, B. (2017). How mergers have driven the consolidation of airline industry. Dallas Business Journal. Retrieved from <https://www.bizjournals.com/dallas/news/2017/04/06/how-mergers-have-driven-consolidation-of-the.html>

[3] Bureau of Transportation Statistics. (n.d.) Database Name: Airline On-Time Performance Data. United States Department of Transportation. Retrieved from <https://www.transtats.bts.gov/Fields.asp>

[4] Enigma Public. (2014). On-time performance dataset for 2008. Retrieved from <https://public.enigma.com/datasets/on-time-performance-2008/0e506417-3449-4051-accb-83c743e6e18d>

[5] Bureau of Transportation Statistics. (n.d.) Airline On-Time Statistics and Delay Causes. United States Department of Transportation. Retrieved from <https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp>

[6] U.S. Department of Transportation. (n.d.) Air Traveler Consumer Report for 2008. Retrieved from <https://www.transportation.gov/airconsumer/air-travel-consumer-reports-2008>