Lister Kom

U19379791

MIT805 Assignment 1

September 12, 2022

1. Introduction

Data like this is crucial in the aviation business, especially when it comes to understanding the causes of flight delays. Flight delays are now universally understood to result in monetary losses for the aviation industry. Over 20% of US flights were delayed in 2018, according to data from the United States Bureau of Transportation Statistics (BTS), resulting in a significant economic effect of around 41 billion US$. Airlines and their passengers are both inconvenienced by these delays. This can ruin airlines' reputations and reduce passenger demand [1]. This dataset will be used in a predictive model, using supervised machine learning algorithms, to predict flight on-time performance, specifically whether a flight will be delayed or not. Such a model can be useful, not only to passengers, but also airlines, for planning purposes. Consequently, losses can be minimised or mitigated.

2. Data

2.1.Data collection, loading and dataset characteristics

The dataset was sourced from Bureau of Transportation Statistics website [2]. The information includes arrival and departure times and flight details for every commercial flight operating within the United States, from October 1995 to 2022. We limited our analysis to the last 5 years due to the large quantity (**volume**) of the dataset i.e. Jan 2018 – June 2022, with the total size of the csv files adding up to 3.46G. The dataset was delivered in the form of 54 zipped csv files organized by month and year. It took about 1 hour and 45 minutes to download the zipped files individually for each month, as there is no date range selector facility on the website. The files were unzipped and placed in the local drive and Google drive. I used Python to extract and merge the files in Jupyterlab. After merging the csv files, the resulting dataset contained about 28 million records, and takes up to 12.4G+ memory in Python. Every year, approximately 7 million flights are recorded from 23 carriers and 325 departure airports. The dataset also came with carrier names csv file, which we joined to the main dataset, using the carrier code. Every flight that occurred during that time period is represented by a row in the dataset, and each column has detailed information on every flight, including the airline, flight date, departure delay, and arrival delay, etc. More details about each variable can be seen in the data dictionary further down [3]. The time it took to load the dataset into Python on a machine with 16GB of RAM was 5 seconds. It took another 2 minutes to concatenate the data. The concatenating process was done in steps, by using a for loop, rather than merging all files at once, which would have obliterated the memory.

Data variety refers to the measure of the richness of the data e.g. including text, graphics, video, and audio. From an analytical standpoint, it is perhaps the greatest barrier to successfully utilising massive volumes of data. Significant obstacles such as incompatible data formats, misaligned data architectures, and inconsistent data semantics can lead to analytic challenges. The airline dataset only consists of text data, however it consists of various variable types (float64, int64 and object). A variety of data types is also a form of data variety. To reduce processing time and RAM utilisation, all the float64 and int64 variable types were changed to float32 and int32 respectively. With regards to data structure, a number of columns contain unstructured data i.e. data that cannot be read by the machine, thus these were converted into suitable format, during the preprocessing phase [4].

Data Velocity Data velocity quantifies the rate at which data is created, streamed, and aggregated. The airline data is reported monthly rather than in real time. The main issue is that there are far too many flights to make streaming data possible. According to the Bureau of Transportation Statistics (from which the data was derived), by 2012, US carriers and international airlines flying into or out of the US had already completed more than 1.37 million flights, equating to over 3,700 flights per day. All those aircraft data recorders would overburden satellite communications. A system like this would be astronomically expensive for airlines, who operate on tiny margins. Real-time tracking of every plan is unnecessary. [5].

Veracity of data relates to its quality. The quality of data is determined by several factors, including where the data was collected, how it was collected, and how it will be analyzed. Veracity determines how trustworthy and significant the data is. Low veracity data typically contains a high percentage of non-valuable, 'noisy,' and meaningless data, which will not aid the analysis. High-quality data, on the other hand, contains a large number of records that are useful for analysis, contributing significantly to the overall results [6]. Our data was collected from the Bureau of Transportation, an organization which prides itself on the data quality provided. The data needs to be reliable and consistent as it is used by the airlines across USA, thus the data has high veracity. Although, there will be 'meaningless' columns deleted, but this is due the fact that they won't be used in this analysis. Another analytic study could potentially still use those fields.

1.2. Metadata (Data Dictionary)

There are 33 variables in the dataset, mixed between numeric and categorical types.

Glossary

Variable	Description	Variable Type
YEAR	The year when the flight occurred (1990 - 1996)	Categorical
MONTH	1 (January) - 12 (December)	Categorical
DAY_OF_MONTH	1 - 31	Categorical
OP_UNIQUE_CARRIER	Unique code assigned by IATA and used to identify a carrier	Categorical
DAY_OF_WEEK	1 (Monday) - 7 (Sunday)	Categorical
ORIGIN_AIRPORT_ID	Departure airport identifier	Categorical
ORIGIN	Origin airport code	Categorical
DEST_AIRPORT_ID	Destination airport identifier	Categorical
DEST	Destination airport code	Categorical
CRS_DEP_TIME	Scheduled departure time e.g. 1630 means 16h30	Numeric
DEP_TIME	Actual departure time e.g. 1630 means 16h30	Numeric
DEP_DELAY	Departure delay (in minutes)	Numeric
TAXI_OUT	Taxi out time (in minutes)	Numeric
TAXI_IN	Taxi in time (in minutes)	Numeric
CRS_ARR_TIME	Scheduled arrival time e.g. 1630 means 16h30	Numeric
ARR_TIME	Actual arrival time e.g. 1630 means 16h30	Numeric
ARR_DELAY	Arrival delay (in minutes)	Numeric
CANCELLED	Was the flight cancelled? (1 => cancelled)	Categorical
CANCELLATION_CODE	Reason for cancellation (A = airline/carrier, B = weather, C = NAS, D = security)	Categorical
DIVERTED	Aircraft landed on different airport that the one scheduled (1 = yes, 0 = no)	Categorical
CRS_ELAPSED_TIME	Scheduled duration of the flight (in minutes)	Numeric
ACTUAL_ELAPSED_TIME	Actual duration of the flight (in minutes)	Numeric
AIR_TIME	The time (in minutes) duration between wheels off (time point that the aircraft's wheels leave the ground) and wheels on time (time point that the aircraft's wheels touch on the ground)	Numeric
FLIGHTS	Number of flights	Categorical
DISTANCE	Distance between airports (miles)	Numeric
CARRIER_DELAY	How long was the flight delayed for, due to the carrier (in minutes) e.g. aircraft cleaning and aircraft damage.	Numeric
WEATHER_DELAY	How long was the weather-related flight delay (in minutes)	Numeric
NAS_DELAY	How long was the flight delayed for, due to NAS (in minutes) e.g. heavy traffic volume.	Numeric
SECURITY_DELAY	How long was the flight delayed for, due to security (in minutes)e.g. evacuation of a terminal	Numeric
LATE_AIRCRAFT_DELAY	How long was the flight delayed for, due to the late arrival of the same aircraft at a previous airport (in minutes)	Numeric
DIV_REACHED_DEST	Diverted flight reaching scheduled destination indicator (1=Yes)	Categorical
DIV_ACTUAL_ELAPSED_TIME	Elapsed time of diverted flight reaching scheduled destination (in minutes).	Numeric
DIV_ARR_DELAY	Difference between scheduled and actual arrival time for a diverted flight reaching scheduled destination (in minutes)	Numeric

Source: [2]

2. Pre-Processing and Explanatory Data Analysis

2.1. Missing Values

Table 1 shows the percentage of missing values per variable. It may appear that a number of variables have large proportion of missing values e.g. 'CARRIER_DELAY', however it is those variables which are populated when an event has occurred e.g. a flight has been delayed. If there was no delayal, these will not be populated. The rows that are populated still therefore provide important information. All the variables associated with departure or arrival e.g. 'DEP_TIME' will also have missing values, where flights have been cancelled. The data is reasonably clean and missing values are indeed justifiably valid.

The variable, 'FLIGHTS' will be deleted as it adds no value, all the rows are populated as 1. Furthermore, given that delay times have been determined already, we will delete the "scheduled" variables ('CRS_DEP_TIME', 'CRS_ARR_TIME','CRS_ELAPSED_TIME','TAXI_OUT' and 'TAXI_IN'). The next variables to be removed are 'ORIGIN' and 'DEST', as they are representing the same information as 'ORIGIN_AIRPORT_ID' and 'DEST_AIRPORT_ID' respectively. For the rest of the variables with a large number of missing values, although justifiably so, for the sake of a simpler model, they will be removed in the base model. These variables are 'CANCELLATION_CODE', 'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY', 'DIV_REACHED_DEST', 'DIV_ARR_DELAY' and 'DIV_ACTUAL_ELAPSED_TIME'. This will leave the dataset with 17 variables remaining.

Table 1: Percentage of missing rows per variable

YEAR	0.000000
MONTH	0.000000
DAY_OF_MONTH	0.000000
DAY_OF_WEEK	0.000000
OP_UNIQUE_CARRIER	0.000000
ORIGIN_AIRPORT_ID	0.000000
ORIGIN	0.000000
DEST_AIRPORT_ID	0.000000
DEST	0.000000
CRS_DEP_TIME	0.000000
DEP_TIME	2.627948
DEP_DELAY	2.633042
TAXI_OUT	2.693205
TAXI_IN	2.736860
CRS_ARR_TIME	0.000000
ARR_TIME	2.711974
ARR_DELAY	2.914566
CANCELLED	0.000000
CANCELLATION_CODE	97.319195
DIVERTED	0.000000
CRS_ELAPSED_TIME	0.000079
ACTUAL_ELAPSED_TIME	2.912616
AIR_TIME	2.937351
FLIGHTS	0.000000
DISTANCE	0.000000
CARRIER_DELAY	82.959495
WEATHER_DELAY	82.959506
NAS_DELAY	82.959506
SECURITY_DELAY	82.959506
LATE_AIRCRAFT_DELAY	82.959506
DIV_REACHED_DEST	99.768265
DIV_ACTUAL_ELAPSED_TIME	99.799598
DIV_ARR_DELAY	99.799391

2.2. Duplicates

The data set contained 14 duplicated records, thus these rows were be deleted from the dataset.

2.3. Summary statistics

After checking for missing values, we begin to evaluate each variable and assess its quality. To do this we assess the summary statistics table.

- We note that maximum values of 'DEP_TIME' (actual departure time) and 'ARR_TIME' (actual arrival time) are 2400 (24h00).
- We note that 'ACTUAL_ELAPSED_TIME' and 'AIR_TIME' have maximum values of 1604 minutes (approx. 27 hours) and 1557 minutes (approx. 26 hours) respectively, which is realistic, as some long-haul flights can be that long.
- We note that "elapsed-time variables" have minimum values below 0, which doesn't make sense. These will be considered as anomalies, thus we decided to delete the rows with such values.
- Average departure and arrival delay is 9.2 and 3.5 minutes respectively.
- Average distance travelled is 788 miles. The minimum distance is 16 miles, which is interestingly short. This could be perhaps a flight that was recalled shortly after it took off.

	count	mean	std	min	25%	50%	75%	max
YEAR	27992263.00000	2019.75518	1.31509	2018.00000	2019.00000	2020.00000	2021.00000	2022.00000
MONTH	27992263.00000	6.30430	3.48790	1.00000	3.00000	6.00000	9.00000	12.00000
DAY_OF_MONTH	27992263.00000	15.74488	8.77568	1.00000	8.00000	16.00000	23.00000	31.00000
DAY_OF_WEEK	27992263.00000	3.96953	2.00179	1.00000	2.00000	4.00000	6.00000	7.00000
ORIGIN_AIRPORT_ID	27992263.00000	12674.76844	1526.77632	10135.00000	11292.00000	12889.00000	14057.00000	16869.00000
DEST_AIRPORT_ID	27992263.00000	12674.75249	1526.77835	10135.00000	11292.00000	12889.00000	14057.00000	16869.00000
CRS_DEP_TIME	27992263.00000	1326.21532	483.03576	1.00000	916.00000	1320.00000	1730.00000	2359.00000
DEP_TIME	27256641.00000	1329.26493	495.57456	1.00000	919.00000	1323.00000	1736.00000	2400.00000
DEP_DELAY	27255215.00000	9.20404	46.82668	-1280.00000	-6.00000	-3.00000	5.00000	3890.00000
TAXI_OUT	27238374.00000	16.64028	9.34474	0.00000	11.00000	14.00000	19.00000	1394.00000
TAXI_IN	27226154.00000	7.50867	6.06978	0.00000	4.00000	6.00000	9.00000	316.00000
CRS_ARR_TIME	27992263.00000	1489.64880	508.15544	1.00000	1108.00000	1515.00000	1915.00000	2400.00000
ARR_TIME	27233120.00000	1468.72332	526.31633	1.00000	1055.00000	1505.00000	1911.00000	2400.00000
ARR_DELAY	27176410.00000	3.45997	48.84418	-1290.00000	-16.00000	-7.00000	6.00000	3864.00000
CANCELLED	27992263.00000	0.02681	0.16152	0.00000	0.00000	0.00000	0.00000	1.00000
DIVERTED	27992263.00000	0.00232	0.04808	0.00000	0.00000	0.00000	0.00000	1.00000
CRS_ELAPSED_TIME	27992241.00000	139.68335	71.06394	-292.00000	89.00000	122.00000	170.00000	1645.00000
ACTUAL_ELAPSED_TIME	27176956.00000	134.17506	71.07230	-1228.00000	82.00000	117.00000	165.00000	1604.00000
AIR_TIME	27170032.00000	110.06371	69.31017	-1244.00000	60.00000	92.00000	140.00000	1557.00000
FLIGHTS	27992263.00000	1.00000	0.00000	1.00000	1.00000	1.00000	1.00000	1.00000
DISTANCE	27992263.00000	788.03063	582.90190	16.00000	362.00000	632.00000	1024.00000	5812.00000
CARRIER_DELAY	4770023.00000	23.23035	68.38429	0.00000	0.00000	2.00000	21.00000	3864.00000
WEATHER_DELAY	4770020.00000	3.82224	31.91340	0.00000	0.00000	0.00000	0.00000	2900.00000
NAS_DELAY	4770020.00000	14.25353	34.65847	0.00000	0.00000	1.00000	18.00000	1741.00000
SECURITY_DELAY	4770020.00000	0.12943	3.51781	0.00000	0.00000	0.00000	0.00000	1245.00000
LATE_AIRCRAFT_DELAY	4770020.00000	25.33759	53.62621	0.00000	0.00000	0.00000	30.00000	2962.00000
DIV_REACHED_DEST	64868.00000	0.86568	0.34100	0.00000	1.00000	1.00000	1.00000	1.00000
DIV_ACTUAL_ELAPSED_TIME	56097.00000	392.43289	224.86690	-139.00000	262.00000	330.00000	435.00000	2414.00000
DIV_ARR_DELAY	56155.00000	253.26511	228.12972	-236.00000	130.00000	181.00000	268.00000	2524.00000

2.3.1. Flights and Departures

Southwest Airlines Co. has the most flights, accounting for 19.0% of all flights, followed by Delta Air Lines Co. at 11.5%. There is a noticeable gap amongst the carriers. Southwest Carriers Co., for instance, accounts for around 19% of flights, which is comparable to the amount of flights chartered by the 16 smallest airlines. The airport with the most departures is Hartsfield-Jackson Atlanta International, which accounts for 4.8% (just over 1.4 million) of all departures, followed by Chicago O'Hare International, which accounts for 4.5% of all departures.

2.3.2. Cancellations

Over the period, 2.7% of the flights were cancelled. Year-on-year, the percentage has been volatile from 1.6% (2018) to 6.1% (2020). The significant increase in cancellations in 2020 was due to the covid pandemic, whereby airlines were not allowed to operate. In terms of months, April has the highest cancellations at 6.9%, which could possibly be distorted by April 2020, where the pandemic was rife. In terms of days, there is no particular day that stands out. Peninsula Airways has the highest cancellation rate at 3.8%. Mammoth Lakes Airport has the highest cancellation rate at 15.4%.

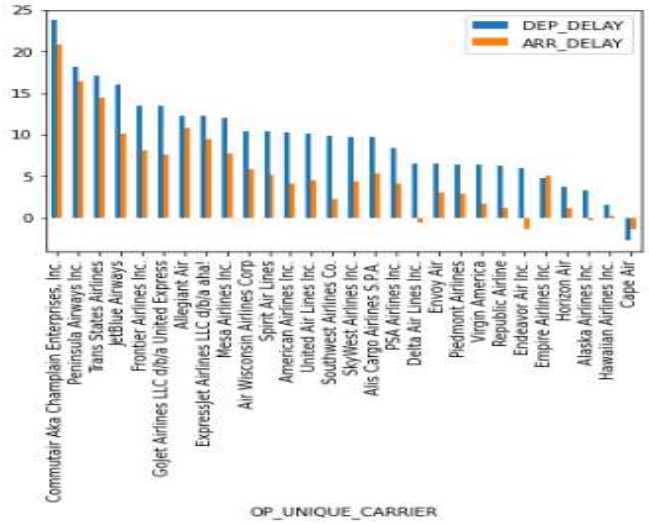
2.3.3. Diverted Flights

Over the period, 0.2% of the flights were diverted. Year-on-year, the percentage has been steady around 0.2%. In terms of months, July has the highest cancellation at 0.3%. In terms of days, there is no particular day that stands out. Peninsula Airways has the highest diversion rate at 2.1%.

2.3.4. Flight Delays

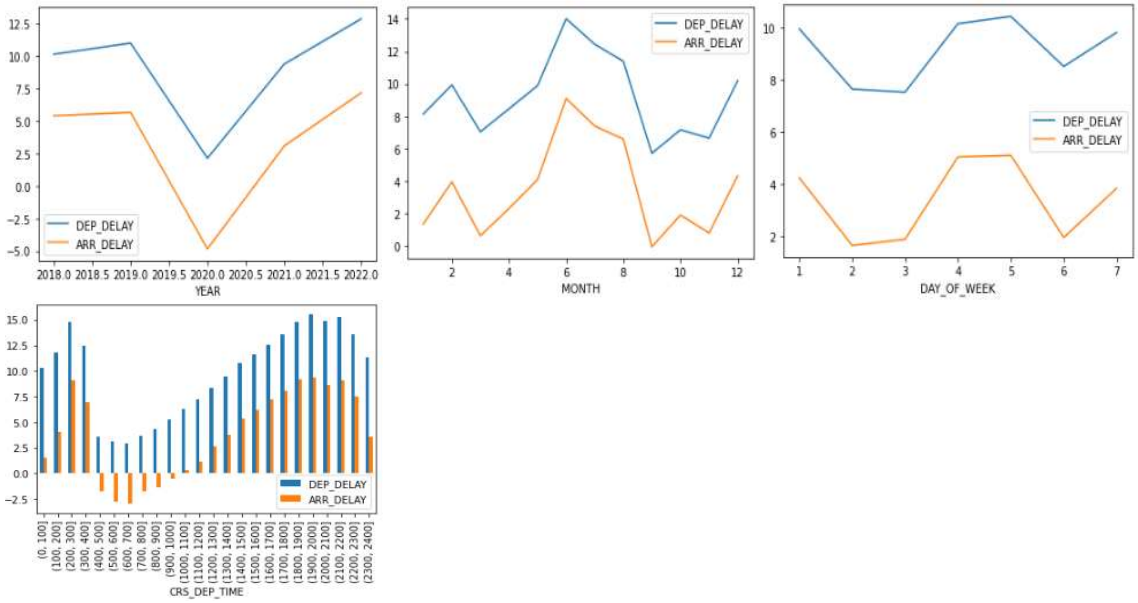
From [Fig 1] we can see it's Commutair which has the longest departure delays, about 24 minutes. Arrival delays are also the longest at about 21 minutes. Cape Air has the shortest departure delays, about 3 minute early. With regards to arrivals, once again, Cape Air, has the shortest arrival delays, about 1 minute early. Generally, airlines have longer departure delays than arrival delays. Arrival delay is probably the most important amongst the two variables though, as a flight can be delayed in departure, but can still arrive at scheduled times. What was noticeable, is that despite Southwest Airline Co, having the most flights, it doesn't have the highest delays.

Fig 1



Looking at departure and arrival delays by year, both have been increasing since 2018, except during the pandemic year. Looking at the delays by month, June appears to be the worst, most likely because it is a busy month of the year as people travel for vacation. Regarding the days, Friday has the delay percentage (just over 10%). Once again, this is probably due to traveling that occurs as the weekend begins. The majority of flights take place in January (9.8%) and March (9.5%), which have lower average delay rates than the other months. Typically though, we would expect that as flight frequencies increase, so will the number of delays. The month of September has the fewest delays. The best time to fly is early in the morning (5am-6am), with delays increasing throughout the day and peaking between 7pm and 10pm.

Fig 2

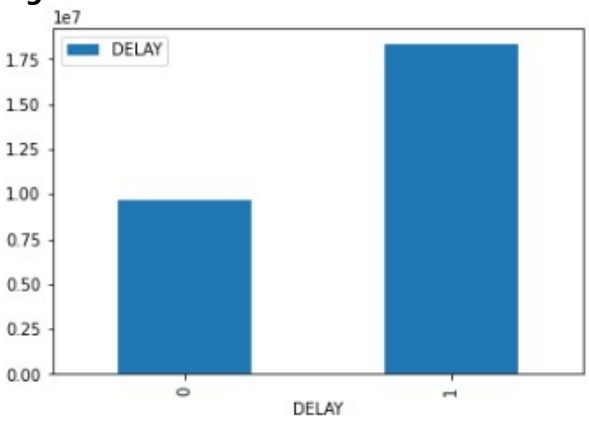


The inspiration for the code and graphs came from [7], [8], [9] and [10].

2.4. Target Variable

A new variable, 'DELAYED' was created, which will be an indicator whether a flight was delayed or not. A flight will be considered delayed if arrival delay time is greater than 0. Fig 3 demonstrates how unbalanced this dataset is with a nearly 2:1 ratio. This must be taken into consideration when evaluating the performance of the models; accuracy alone will not suffice, therefore I will also include Precision and Recall [11].

Fig 3



Using the target variable, it can seen that it varies significantly by carriers, between 33.9% and 75.3%. Endeavor Air Inc has the highest percentage of delayed flights at 75.3%. The percentage of delays doesn't vary much by the months of the year, with June having the lowest percentage (60%) and September having the highest percentage (71.5%). Similarly across the days of the week, there is not much variation. It varies quite significantly by airport, with Victoria Airport having the highest percentage delays (85%). The only airport without any delays is Youngstown/Warren Airport (Fig 4).

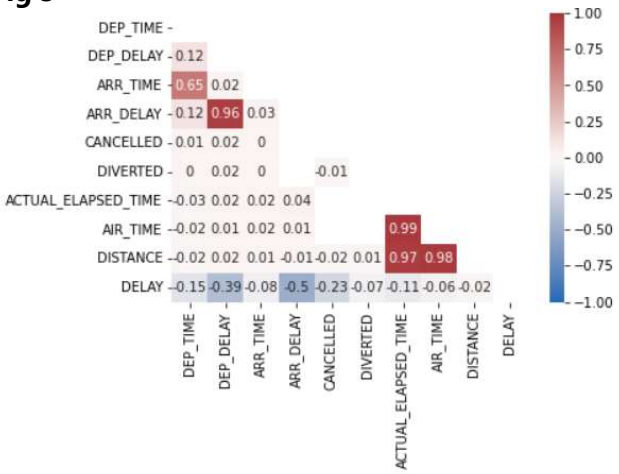
Fig 4

```
ORIGIN_AIRPORT_ID
Victoria, TX: Victoria Regional                                    85.45%
Fort Leonard Wood, MO: Waynesville-St. Robert Regional Forney Field  83.52%
Fort Dodge, IA: Fort Dodge Regional                                81.85%
Hibbing, MN: Range Regional                                        81.15%
Pocatello, ID: Pocatello Regional                                  81.12%
                                                                    ...
Adak Island, AK: Adak                                              40.18%
Pago Pago, TT: Pago Pago International                             39.18%
Wilmington, DE: New Castle                                         38.79%
Cold Bay, AK: Cold Bay Airport                                     28.63%
Youngstown/Warren, OH: Youngstown-Warren Regional                   0.00%
```

2.5. Correlations

For this dataset, I would predict a high correlation between departure delay and arrival delay i.e. the longer the airline is delayed at departure, it's expected that the longer the airline would be delayed at arrival. The actual correlation is 96%. This points towards multicollinearity for these variables, we most likely need just one of them. Against the target variable ('DELAY'), I would also predict high correlations vs airline, airport, month of the year. Against the days of the week, I would expect weak correlation. The correlation below (Fig 5) depicts some variables which are highly correlated, thus posing a multi collinearity scenario.

Fig 5


DEP_TIME
DEP_DELAY
ARR_TIME
ARR_DELAY
CANCELLED
DIVERTED
ACTUAL_ELAPSED_TIME
AIR_TIME
DISTANCE
DELAY
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

References

[1] JaHerbas, "GitHub: JaHerbas/Predicting Flight Delays," [Online]. Available: https://github.com/JaHerbas/Predicting_Flight_Delays. [Accessed 28 August 2022].

[2] "Bureau of Statistics," Bureau of Statistics, [Online]. Available: https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr. [Accessed 1 September 2022].

[3] "Data Expo 2009: Airline on time data," [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7. [Accessed 29 August 2022].

[4] J. Spacey, "Simplicable," [Online]. Available: https://simplicable.com/new/data-variety#:~:text=Data%20variety%20is%20the%20diversity%20of%20data%20in,such%20as%20different%20types%20of%20database%20or%20file.. [Accessed 8 September 2022].

[5] J. Golson, "Wired," [Online]. Available: https://www.wired.com/2015/01/why-we-dont-need-real-time-flight-tracking/. [Accessed 29 August 2022].

[6] I. Team, "Indicative," [Online]. Available: https://www.indicative.com/resource/data-veracity/. [Accessed 29 August 2022].

[7] S. P. Enmhe, "GitHub: Shawnemhe," [Online]. Available: https://shawnemhe.github.io/udacity-data-analyst/p6/python_eda/python_eda.html. [Accessed 8 September 2022].

[8] J. Brooks, "Kaggle: Airlines Delay and Cancellation Analysis," [Online]. Available: https://www.kaggle.com/code/jcbrooks/airlines-delay-and-cancellation-analysis. [Accessed 2 September 2022].

[9] Y. Wang, "Medium: Predicting Flight Delays through modeling U.S. Flight Data," [Online]. Available: https://medium.com/analytics-vidhya/modeling-flight-delays-through-u-s-flight-data-2f0b3d7e2c89. [Accessed 29 August 2022].

[10] S. Yıldırım, "TowardsDataScience: A Practical Guide for Exploratory Data Analysis: Flight Delays," [Online]. Available: https://towardsdatascience.com/a-practical-guide-for-exploratory-data-analysis-flight-delays-f8a713ef7121. [Accessed 28 August 2022].

[11] J. Herbas, "Medium," [Online]. Available: https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c. [Accessed 9 Septemember 2022].

[12] A. Vera, "Kaggle: Flight Delay EDA (Exploratory Data Analysis)," [Online]. Available: https://www.kaggle.com/code/adveros/flight-delay-eda-exploratory-data-analysis/notebook. [Accessed 29 August 2022].