Getting the Data

Jakob Bjerre Eriksen, Keyla Jaylin Barcenas, Keval Darshan Amin, Noa Adriana Gonzalez

Thursday 18th December, 2025

This notebook extracts data from Kaggle and preprocesses it to contain only the relevant data, and get rid of missings.

Furthermore, simple overviews of the raw and filtered data are shown.

```
import pandas as pd
import numpy as np
from helper_functions import get_full_data, summarize_df
```

1 Extracting Data from Kaggle

```
# Download data into cache
data_path = get_full_data()
```

```
# Create pandas dataframes of the datasets
airlines_df = pd.read_csv(data_path+"/airlines.csv")
airports_df = pd.read_csv(data_path+"/airports.csv")
flights_full_df = pd.read_csv(data_path+"/flights.csv", low_memory=False)
```

2 Overviews and Summaries of the Different Raw Datasets

```
# Overview of the airlines
airlines_df.head()
```

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways

```
# Overview of the airports
airports_df.head()
```

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
0	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
1	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
2	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
3	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
4	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447

```
# Overview of the raw dataset containing the flights
flights_full_df.head()
```

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	...	SCHEDULED_TIME	DISTANCE	DIVERTED	CANCELLED	CANCELLATION_REASON	AIR_SYSTEM_DELAY	SECURITY_DELAY	AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
0	2015	1	1	4	AS	98	N407AS	ANC	SEA	5	...	205.0	1448	0	0	NaN	NaN	NaN	NaN	NaN	NaN
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10	...	280.0	2330	0	0	NaN	NaN	NaN	NaN	NaN	NaN
2	2015	1	1	4	US	840	N171US	SFO	CLT	20	...	286.0	2296	0	0	NaN	NaN	NaN	NaN	NaN	NaN
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	...	285.0	2342	0	0	NaN	NaN	NaN	NaN	NaN	NaN
4	2015	1	1	4	AS	135	N527AS	SEA	ANC	25	...	235.0	1448	0	0	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 31 columns

```
# Summary of the raw dataset containing all the flights
summary_flights = summarize_df(flights_full_df,
                               "figures/df_summary_all_flights.png")
summary_flights
```

	dtype	min	max	mean	std	unique_vals	missing_pct
AIRLINE	object	NaN	NaN	NaN	NaN	14	0.000000
AIRLINE_DELAY	float64	0.0	1971.0	18.969547	48.161642	1067	0.817250
AIR_SYSTEM_DELAY	float64	0.0	1134.0	13.480568	28.003679	570	0.817250
AIR_TIME	float64	7.0	690.0	113.511628	72.230822	675	0.018056
ARRIVAL_DELAY	float64	-87.0	1971.0	4.407057	39.271297	1240	0.018056
ARRIVAL_TIME	float64	1.0	2400.0	1476.491188	526.319737	1440	0.015898
CANCELLATION_REASON	object	NaN	NaN	NaN	NaN	4	0.984554
CANCELLED	int64	0.0	1.0	0.015446	0.123320	2	0.000000
DAY	int64	1.0	31.0	15.704594	8.783425	31	0.000000
DAY_OF_WEEK	int64	1.0	7.0	3.926941	1.988845	7	0.000000
DEPARTURE_DELAY	float64	-82.0	1988.0	9.370158	37.080942	1217	0.014805
DEPARTURE_TIME	float64	1.0	2400.0	1335.204439	496.423260	1440	0.014805
DESTINATION_AIRPORT	object	NaN	NaN	NaN	NaN	629	0.000000
DISTANCE	int64	21.0	4983.0	822.356495	607.784287	1363	0.000000
DIVERTED	int64	0.0	1.0	0.002610	0.051020	2	0.000000
ELAPSED_TIME	float64	14.0	766.0	137.006189	74.211072	712	0.018056
FLIGHT_NUMBER	int64	1.0	9855.0	2173.092742	1757.063999	6952	0.000000
LATE_AIRCRAFT_DELAY	float64	0.0	1331.0	23.472838	43.197018	695	0.817250
MONTH	int64	1.0	12.0	6.524085	3.405137	12	0.000000
ORIGIN_AIRPORT	object	NaN	NaN	NaN	NaN	628	0.000000
SCHEDULED_ARRIVAL	int64	1.0	2400.0	1493.808249	507.164696	1435	0.000000
SCHEDULED_DEPARTURE	int64	1.0	2359.0	1329.602470	483.751821	1321	0.000000
SCHEDULED_TIME	float64	18.0	718.0	141.685892	75.210582	550	0.000001
SECURITY_DELAY	float64	0.0	573.0	0.076154	2.143460	154	0.817250
TAIL_NUMBER	object	NaN	NaN	NaN	NaN	4897	0.002530
TAXI_IN	float64	1.0	248.0	7.434971	5.638548	185	0.015898
TAXI_OUT	float64	1.0	225.0	16.071662	8.895574	184	0.015303
WEATHER_DELAY	float64	0.0	1211.0	2.915290	20.433336	632	0.817250
WHEELS_OFF	float64	1.0	2400.0	1357.170841	498.009356	1440	0.015303
WHEELS_ON	float64	1.0	2400.0	1471.468609	522.187945	1440	0.015898
YEAR	int64	2015.0	2015.0	2015.000000	0.000000	1	0.000000

3 Filtering the Raw Flights Dataset

```
filtered_flights_df = flights_full_df.copy()

# Filtering for flights with airports in California as origin airport
airports_CA = airports_df.loc[airports_df["STATE"] == "CA", "IATA_CODE"].tolist()
filtered_flights_df = filtered_flights_df[
    # (filtered_flights_df["DESTINATION_AIRPORT"].isin(airports_CA)) |
    (filtered_flights_df["ORIGIN_AIRPORT"].isin(airports_CA))
    ]

# Assuming that nan values in AIRLINE_DELAY, AIR_SYSTEM_DELAY, LATE_AIRCRAFT_DELAY,
# SECURITY_DELAY and WEATHER_DELAY means that the flight was not delayed in those
# areas, setting the values of the variables equal to 0
cols_fill_nan = ["AIRLINE_DELAY",
                 "AIR_SYSTEM_DELAY",
                 "LATE_AIRCRAFT_DELAY",
                 "SECURITY_DELAY",
                 "WEATHER_DELAY"]
filtered_flights_df[cols_fill_nan] = filtered_flights_df[cols_fill_nan].fillna(0)

# Dropping specific columns
cols_remove = ["CANCELLATION_REASON", # Because of too many missings
```

```
              "YEAR", # Because all values are 2015
              "DIVERTED" # Because all remaining values are 0
              ]
filtered_flights_df = filtered_flights_df.drop(cols_remove, axis=1)
```

```
# As some of the flights are cancelled, some missings in the data are only missings
# because the plane did not take off, and should thus not be removed
# Removing remaining rows with true missings (they only make up < 0.5% of the total remaining data)
cols_remove_rows_false = ["AIR_TIME",
                  "ARRIVAL_DELAY",
                  "ARRIVAL_TIME",
                  "DEPARTURE_DELAY",
                  "DEPARTURE_TIME",
                  "ELAPSED_TIME",
                  "TAXI_IN",
                  "TAXI_OUT",
                  "WHEELS_OFF",
                  "WHEELS_ON"]
filtered_flights_df = filtered_flights_df[
    (filtered_flights_df["CANCELLED"] == 1) |
    (~filtered_flights_df[cols_remove_rows_false].isna().any(axis=1))
    ]

# Removing remaining rows with missings, but where data should be present no matter if
# the plane took off or not
cols_remove_rows_true = ["TAIL_NUMBER"]
filtered_flights_df = filtered_flights_df.dropna(subset=cols_remove_rows_true)

# Removing rows where the flight is cancelled, but the row still shows data for the trip
filtered_flights_df = filtered_flights_df[
    ~((filtered_flights_df["CANCELLED"] == 1) &
    (filtered_flights_df[cols_remove_rows_false].notna().any(axis=1)))
    ]
```

4 Overview of the Filtered Flights Dataset

```
# Overview of the filtered dataset containing the flights
summary_filtered_flights = summarize_df(filtered_flights_df,
                                        "figures/df_summary_filtered_flights.png")
summary_filtered_flights
```

	dtype	min	max	mean	std	unique_vals	missing_pct
AIRLINE	object	NaN	NaN	NaN	NaN	13	0.000000
AIRLINE_DELAY	float64	0.0	1665.0	3.163064	20.627670	563	0.000000
AIR_SYSTEM_DELAY	float64	0.0	748.0	2.027332	11.235341	336	0.000000
AIR_TIME	float64	8.0	409.0	127.601231	88.060252	391	0.010222
ARRIVAL_DELAY	float64	-69.0	1665.0	5.122734	36.729980	702	0.010222
ARRIVAL_TIME	float64	1.0	2400.0	1474.080986	547.001758	1440	0.010222
CANCELLED	int64	0.0	1.0	0.010222	0.100588	2	0.000000
DAY	int64	1.0	31.0	15.692449	8.776948	31	0.000000
DAY_OF_WEEK	int64	1.0	7.0	3.917096	1.991989	7	0.000000
DEPARTURE_DELAY	float64	-38.0	1670.0	9.457042	35.225836	683	0.010222
DEPARTURE_TIME	float64	1.0	2400.0	1324.310995	520.161601	1375	0.010222
DESTINATION_AIRPORT	object	NaN	NaN	NaN	NaN	96	0.000000
DISTANCE	int64	77.0	2704.0	991.654938	799.059187	331	0.000000
ELAPSED_TIME	float64	23.0	484.0	149.994630	90.379776	420	0.010222
FLIGHT_NUMBER	int64	1.0	9855.0	2115.949736	1908.117498	5917	0.000000
LATE_AIRCRAFT_DELAY	float64	0.0	1102.0	5.104048	21.623389	448	0.000000
MONTH	int64	1.0	12.0	6.221808	3.384171	11	0.000000
ORIGIN_AIRPORT	object	NaN	NaN	NaN	NaN	22	0.000000
SCHEDULED_ARRIVAL	int64	1.0	2359.0	1487.817366	530.778103	1330	0.000000
SCHEDULED_DEPARTURE	int64	1.0	2359.0	1322.168889	508.058394	1188	0.000000
SCHEDULED_TIME	float64	39.0	395.0	153.895688	91.930221	346	0.000000
SECURITY_DELAY	float64	0.0	440.0	0.013097	1.053790	77	0.000000
TAIL_NUMBER	object	NaN	NaN	NaN	NaN	3904	0.000000
TAXI_IN	float64	1.0	248.0	7.259847	5.444749	138	0.010222
TAXI_OUT	float64	1.0	176.0	15.133552	7.170978	132	0.010222
WEATHER_DELAY	float64	0.0	916.0	0.188518	4.956099	230	0.000000
WHEELS_OFF	float64	1.0	2400.0	1338.349737	522.040646	1376	0.010222
WHEELS_ON	float64	1.0	2400.0	1470.315109	542.356793	1440	0.010222

5 Saving the Datasets

```
# Saving the datasets to CSV files in the project folder
airlines_df.to_csv("data/airlines.csv", index=False)
airports_df.to_csv("data/airports.csv", index=False)
filtered_flights_df.to_csv("data/filtered_flights.csv", index=False)
```