Eksploracja danych - etap 1

Krzysztof Nasuta 193328, Filip Dawidowski 193433, Aleks Iwicki 193354

1. Ogólny opis zbioru

Departament Transportu Stanów Zjednoczonych (DOT), za pośrednictwem Biura Statystyki Transportu, monitoruje punktualność krajowych lotów realizowanych przez dużych przewoźników lotniczych. Zbiorcze informacje na temat liczby lotów punktualnych, opóźnionych, odwołanych oraz przekierowanych są publikowane w comiesięcznym raporcie „Air Travel Consumer Report" oraz w tym zestawie danych dotyczącym opóźnień i odwołań lotów z 2015 roku.
Pojedynczy wiersz zbioru danych reprezentuje pojedynczy lot. Zbiór danych zawiera informacje o czasie odlotu i przylotu, czasie opóźnienia, przyczynie opóźnienia, a także inne szczegóły dotyczące lotu.

2. Określenie celu eksploracji i kryteriów sukcesu

Celem eksploracji jest predykcja, czy lot będzie opóźniony, czy nie. Uznajemy, że lot jest opóźniony, jeśli atrybut ARRIVAL_DELAY jest większy niż 15. Dodatkowo chcemy zrozumieć, które czynniki mają największy wpływ na opóźnienia lotów.
Kryterium sukcesu jest osiągnięcie dokładności klasyfikacji opóźnienia na poziomie 85%. W naszym przypadku chcemy skupić się na maksymalizacji czułości modelu ponad swoistość, ponieważ błędne wykrycie opóźnienia nie jest tak istotne, jak przeoczenie rzeczywistego opóźnienia.

3. Charakterystyka zbioru

- Pochodzenie: Kaggle
- Liczba przykładów: 5819080
- Format: CSV (3 pliki: `flights.csv` - właściwy zbiór, `airports.csv` - informacje o lotniskach, `airlines.csv` - informacje o liniach lotniczych)
- Ilość zbiorów danych: 1

4. Opis atrybutów

Nazwa	Typ	Opis
YEAR	Numeryczny	Rok lotu
MONTH	Numeryczny	Miesiąc lotu
DAY	Numeryczny	Dzień miesiąca lotu
DAY_OF_WEEK	Numeryczny	Dzień tygodnia lotu
AIRLINE	Nominalny	Identyfikator linii lotniczej
FLIGHT_NUMBER	Numeryczny	Numer lotu
TAIL_NUMBER	Nominalny	Numer rejestracyjny samolotu
ORIGIN_AIRPORT	Nominalny	Kod IATA lotniska wylotu
DESTINATION_AIRPORT	Nominalny	Kod IATA lotniska przylotu
SCHEDULED_DEPARTURE	Numeryczny	Planowany czas odlotu (w formacie HHMM)
DEPARTURE_TIME	Numeryczny	Czas odlotu (w formacie HHMM)

Nazwa	Typ	Opis
DEPARTURE_DELAY	Numeryczny	Całkowite opóźnienie odlotu (w minutach)
TAXI_OUT	Numeryczny	Ilość minut spędzonych na kołowaniu przed odlotem
WHEELS_OFF	Numeryczny	Czas startu (w formacie HHMM)
SCHEDULED_TIME	Numeryczny	Planowany czas lotu (w minutach)
ELAPSED_TIME	Numeryczny	Całkowity czas lotu (w minutach) = AIR_TIME+TAXI_IN+TAXI_OUT
AIR_TIME	Numeryczny	Czas lotu (w minutach)
DISTANCE	Numeryczny	Dystans lotu (w milach)
WHEELS_ON	Numeryczny	Czas lądowania (w formacie HHMM)
TAXI_IN	Numeryczny	Ilość minut spędzonych na kołowaniu po lądowaniu
SCHEDULED_ARRIVAL	Numeryczny	Planowany czas przylotu (w formacie HHMM)
ARRIVAL_TIME	Numeryczny	Czas przylotu (w formacie HHMM) = WHEELS_ON+TAXI_IN
ARRIVAL_DELAY	Numeryczny	Całkowite opóźnienie przylotu (w minutach) = ARRIVAL_TIME-SCHEDULED_ARRIVAL
DIVERTED	Prawda/Fałsz	Czy lot był przekierowany? (1-tak, 0-nie)
CANCELLED	Prawda/Fałsz	Czy lot był odwołany? (1-tak, 0-nie)
CANCELLATION_REASON	Nominalny	Przyczyna odwołania lotu (A - przewoźnik, B - pogoda, C - National Air System, D - Bezpieczeństwo), tylko dla odwołanych lotów
AIR_SYSTEM_DELAY	Numeryczny	Opóźnienie spowodowane przez system lotniczy (w minutach)
SECURITY_DELAY	Numeryczny	Opóźnienie spowodowane przez kontrole bezpieczeństwa (w minutach)
AIRLINE_DELAY	Numeryczny	Opóźnienie spowodowane przez przewoźnika (w minutach)
LATE_AIRCRAFT_DELAY	Numeryczny	Opóźnienie spowodowane przez samolot (w minutach)
WEATHER_DELAY	Numeryczny	Opóźnienie spowodowane przez pogodę (w minutach)

4.1. Brakujące dane

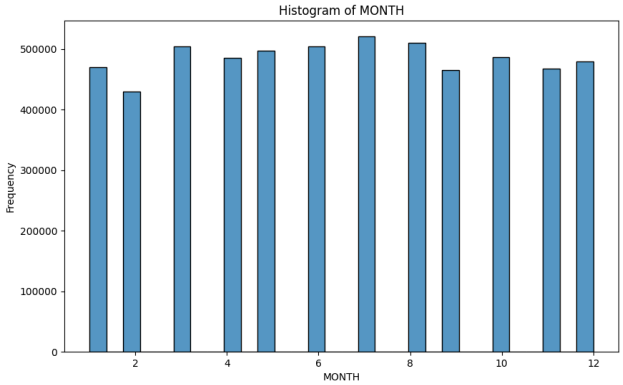
W zbiorze danych występowały nieznaczące braki, lecz stanowiły one nieznaczną część całego zbioru, który jest bardzo duży (ponad 5 milionów przykładów). Z uwagi na ten fakt, zostały one odfiltrowane podczas przetwarzania.

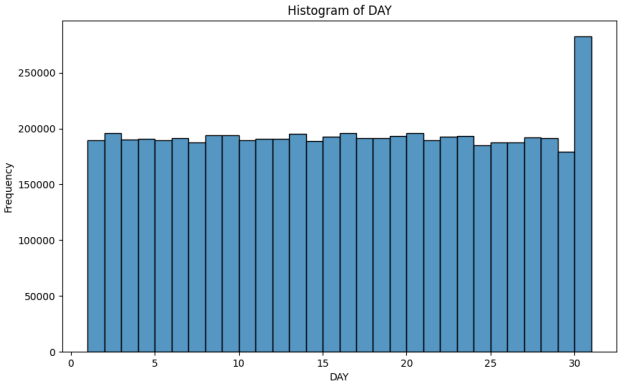
4.2. Dane niezrozumiałe

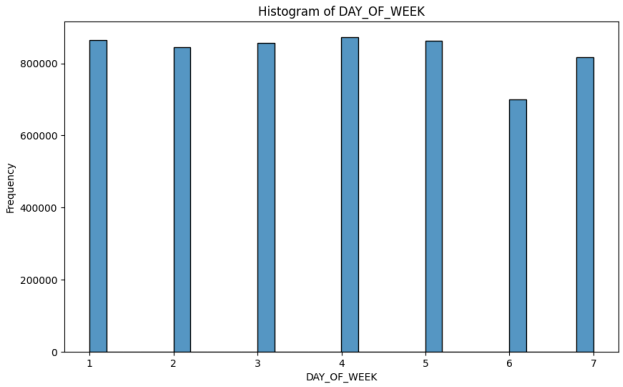
brak

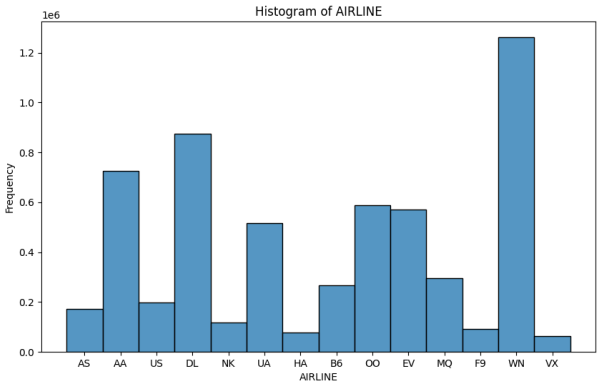
5. Wynik eksploracyjnej analizy danych

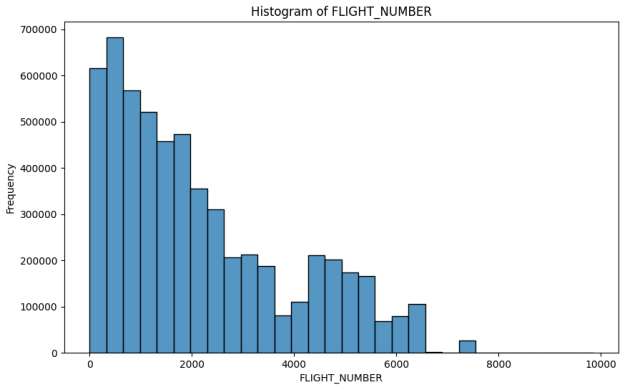
5.1. Rozkłady wartości atrybutów

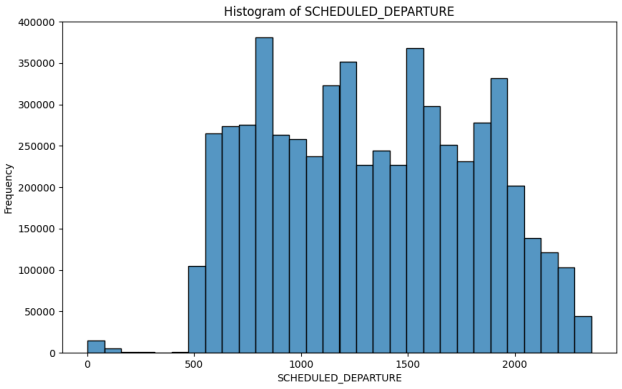


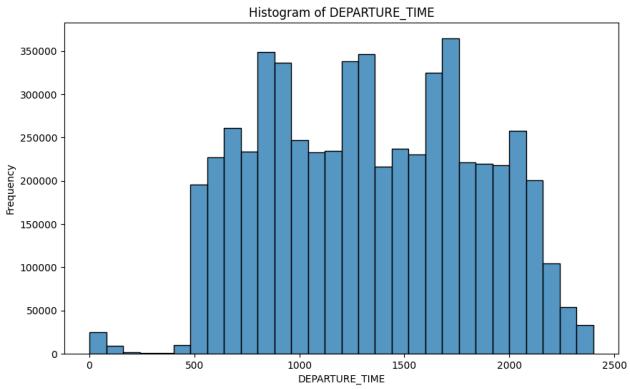


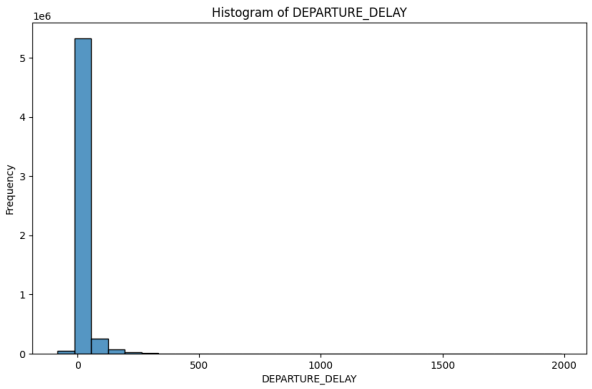


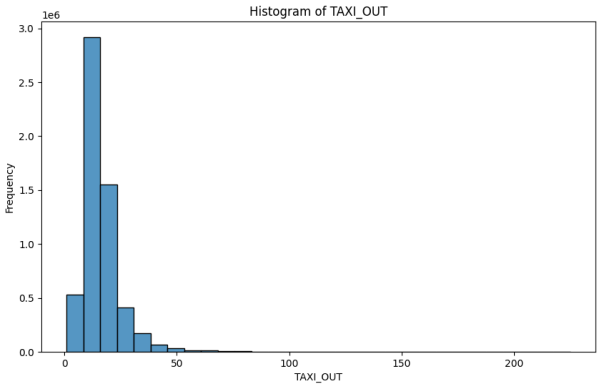


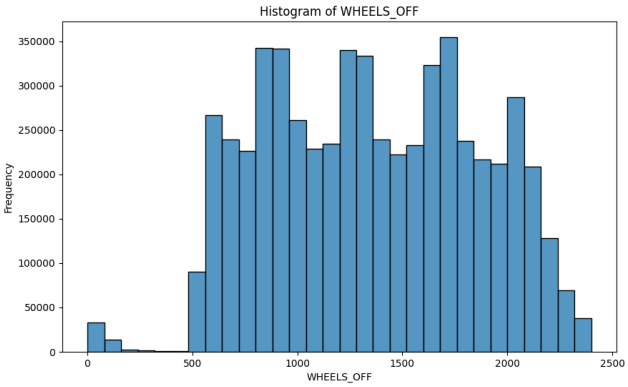


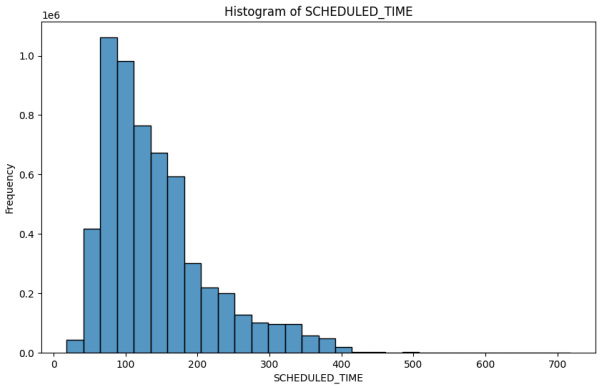


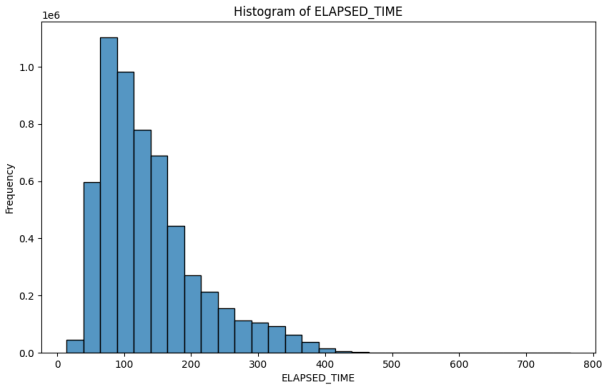


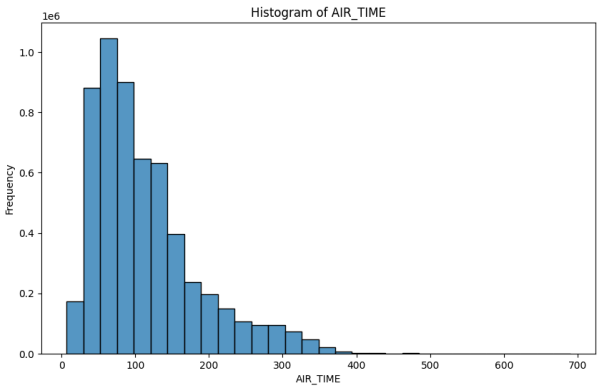


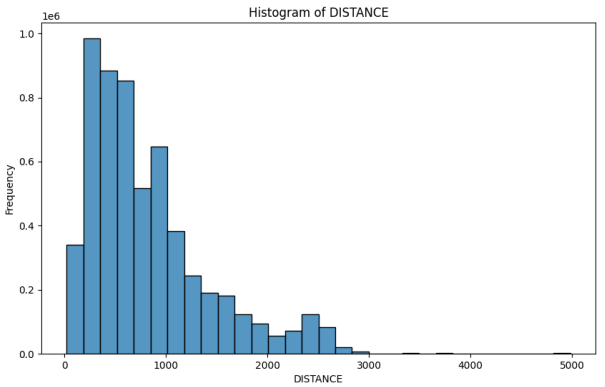

Histogram of TAXI_OUT
1e6
Frequency
TAXI_OUT

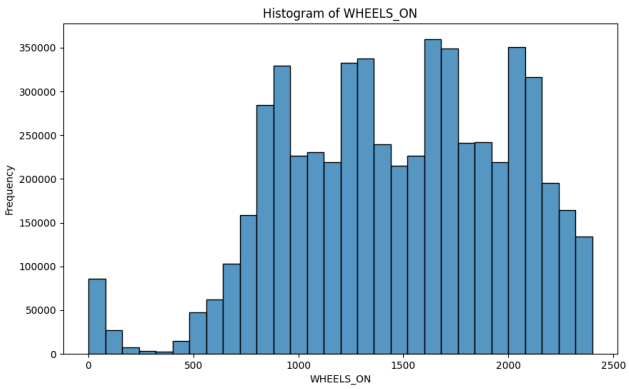

Histogram of WHEELS_OFF
Frequency
WHEELS_OFF

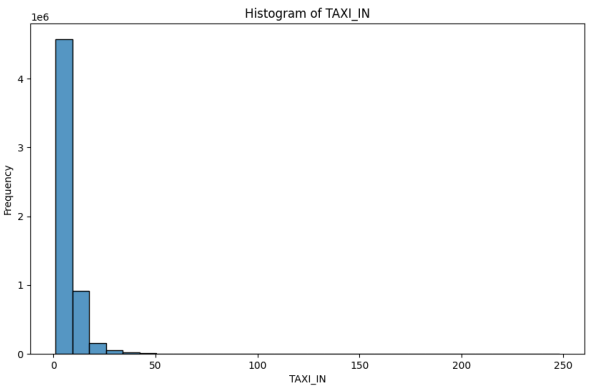

Histogram of SCHEDULED_TIME
1e6
Frequency
SCHEDULED_TIME

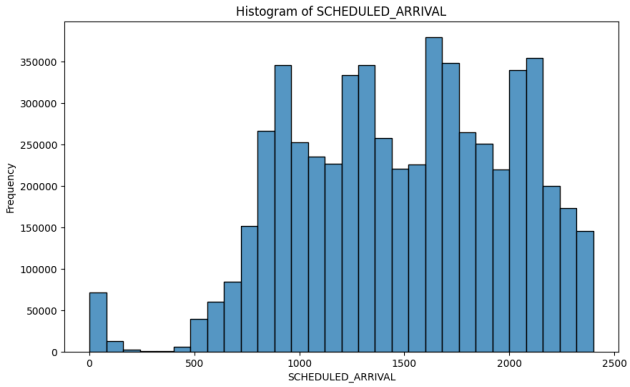

Histogram of ELAPSED_TIME
1e6
Frequency
ELAPSED_TIME

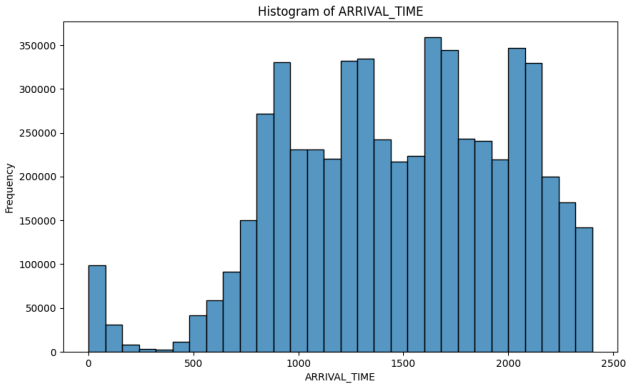

Histogram of AIR_TIME
1e6
Frequency
AIR_TIME

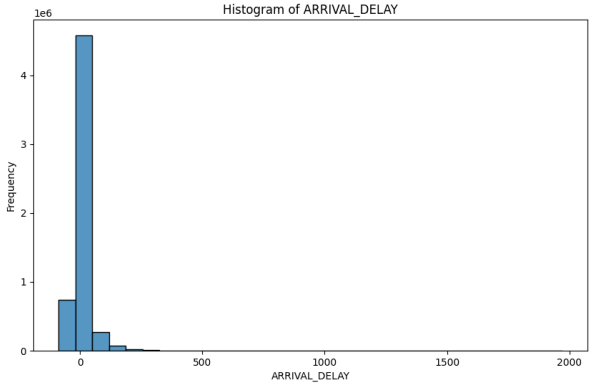

Histogram of DISTANCE
1e6
Frequency
DISTANCE

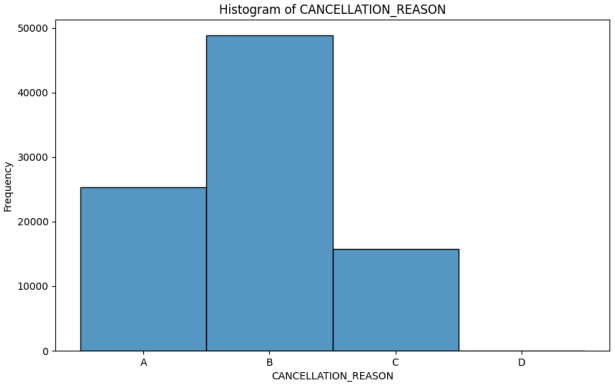

Histogram of WHEELS_ON
Frequency
WHEELS_ON

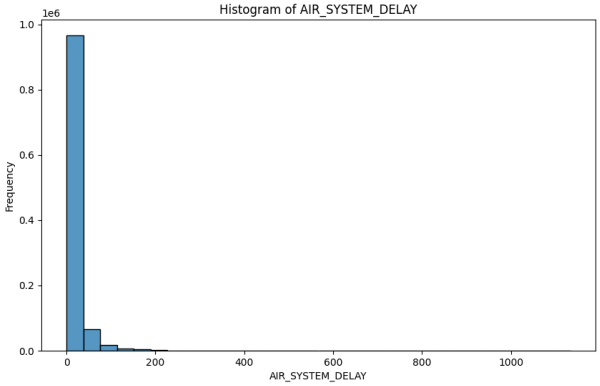

Histogram of TAXI_IN
1e6
Frequency
TAXI_IN

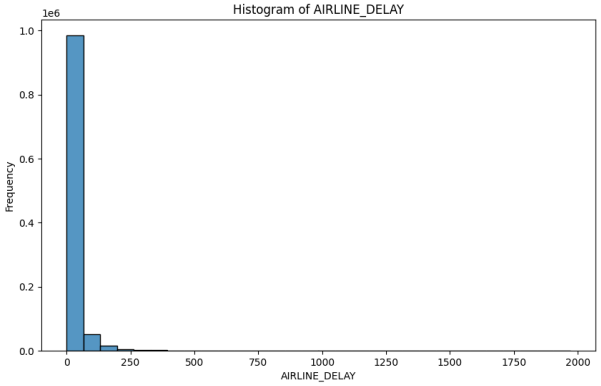


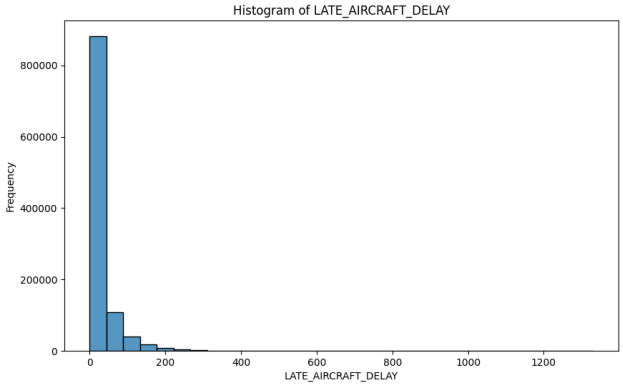


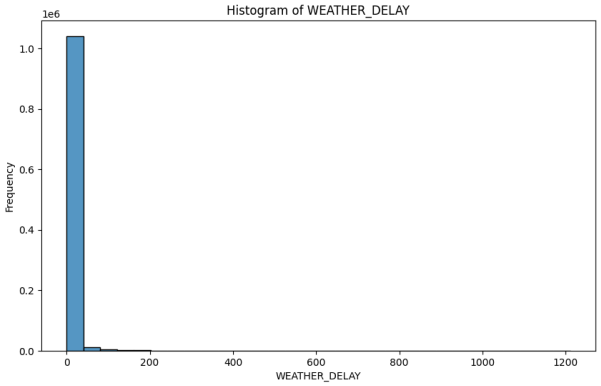








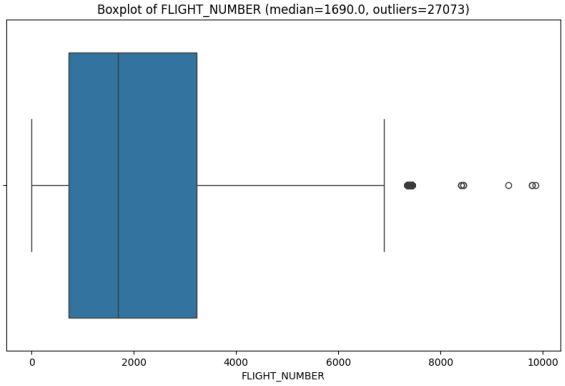


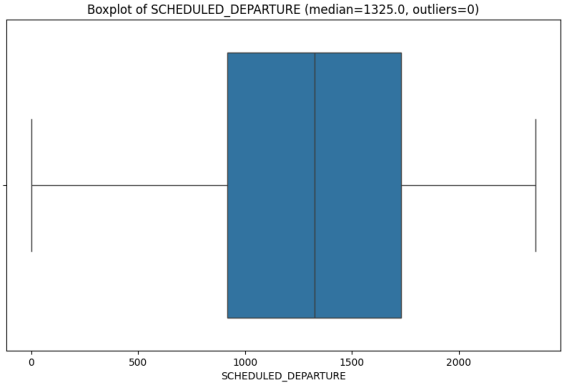


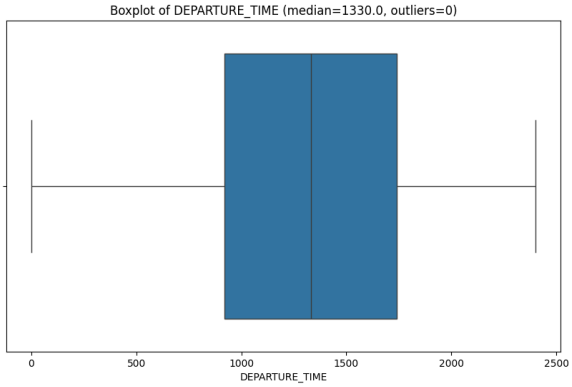
- W przypadku atrybutu ARRIVAL_DELAY zauważalna jest przewaga lotów punktualnych lub z niewielkim opóźnieniem (poniżej 15 minut) względem lotów znacząco opóźnionych. Klasy są niezbalansowane, co może wpłynąć na skuteczność modeli klasyfikacyjnych.

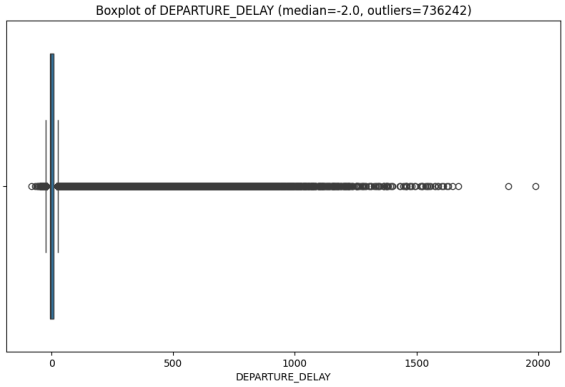
- Rozkłady większości atrybutów numerycznych, takich jak DEPARTURE_DELAY, AIR_TIME, TAXI_OUT czy SCHEDULED_TIME, nie przypominają rozkładu normalnego. Najczęściej obserwujemy rozkład prawoskośny – większość wartości skupia się w niższych przedziałach, a ogon rozkładu jest wydłużony w stronę wyższych wartości.
- Wysokie, rzadko występujące wartości w atrybutach takich jak WEATHER_DELAY, AIRLINE_DELAY oraz LATE_AIRCRAFT_DELAY mogą wskazywać na zdarzenia nietypowe, takie jak intensywne burze, problemy techniczne lub opóźnienia łańcuchowe wynikające z wcześniejszych lotów. Tego typu przypadki mają istotne znaczenie dla analizy przyczyn opóźnień i mogą być kluczowe przy budowie predykcyjnych modeli.
- Wartości atrybutu DISTANCE rozkładają się nierównomiernie –większość lotów odbywa się na krótkich i średnich dystansach, co znajduje odzwierciedlenie w rozkładzie. Długodystansowe loty są mniej liczne.

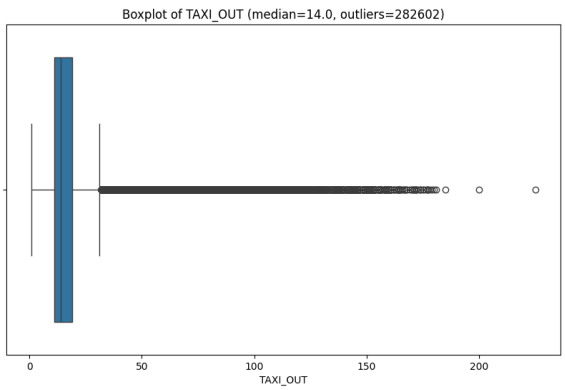
5.2. Punkty oddalone

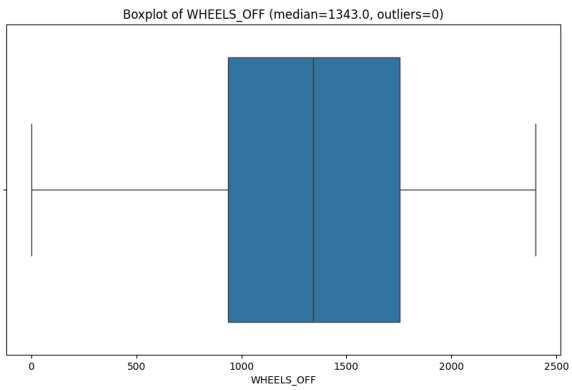


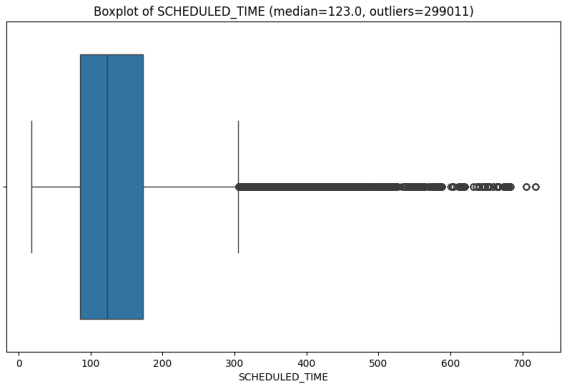


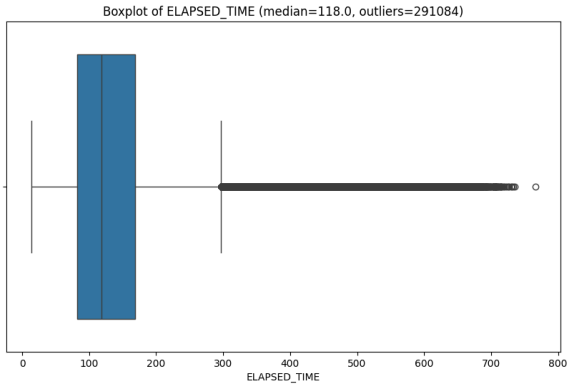


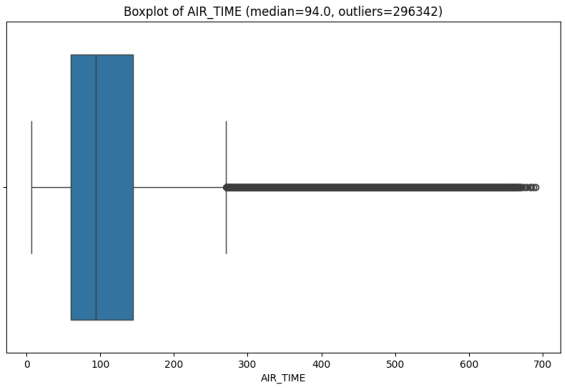


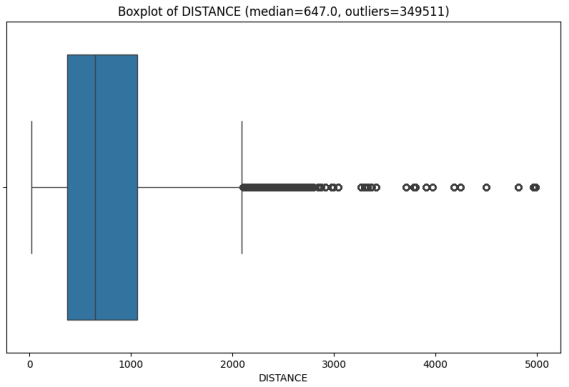


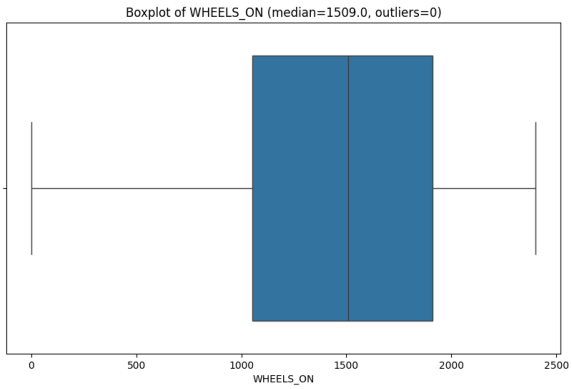


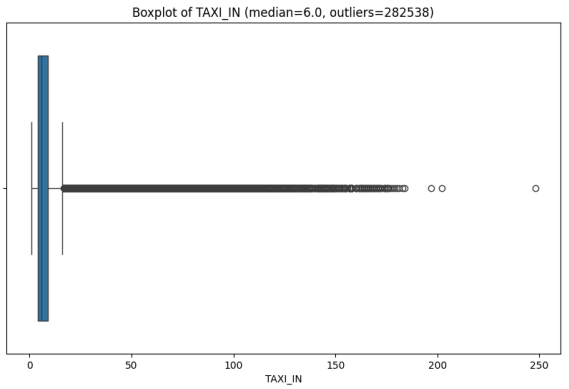

Boxplot of SCHEDULED_TIME (median=123.0, outliers=299011)
0
100
200
300
400
500
600
700
SCHEDULED_TIME

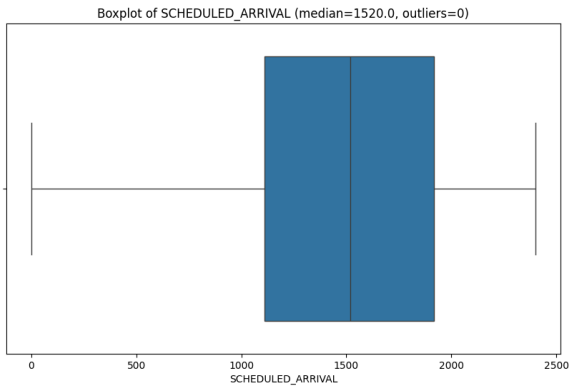

Boxplot of ELAPSED_TIME (median=118.0, outliers=291084)
0
100
200
300
400
500
600
700
800
ELAPSED_TIME

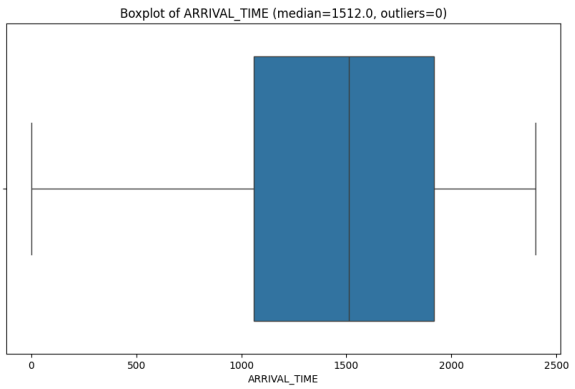

Boxplot of AIR_TIME (median=94.0, outliers=296342)
0
100
200
300
400
500
600
700
AIR_TIME

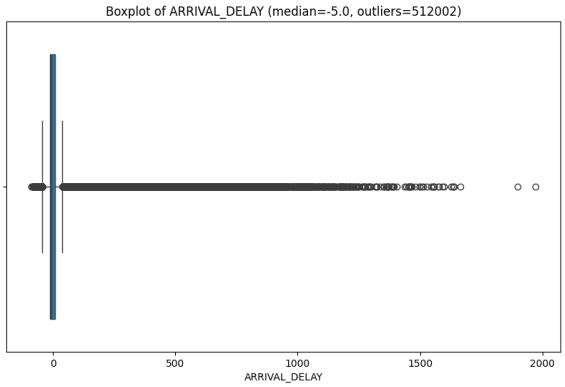

Boxplot of DISTANCE (median=647.0, outliers=349511)
0
1000
2000
3000
4000
5000
DISTANCE

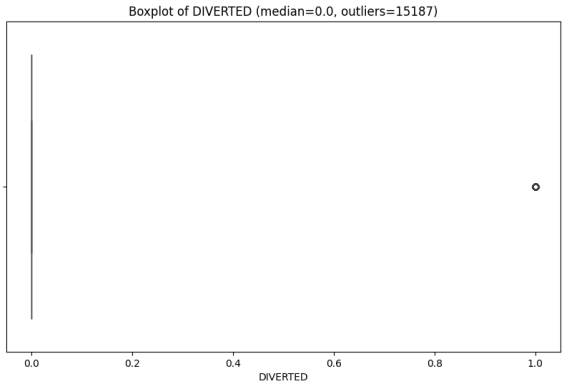

Boxplot of WHEELS_ON (median=1509.0, outliers=0)
0
500
1000
1500
2000
2500
WHEELS_ON

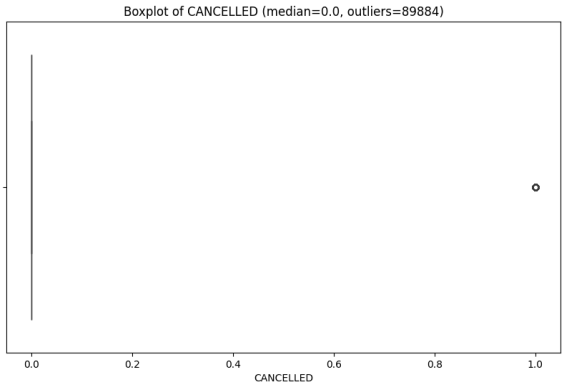

Boxplot of TAXI_IN (median=6.0, outliers=282538)
0
50
100
150
200
250
TAXI_IN

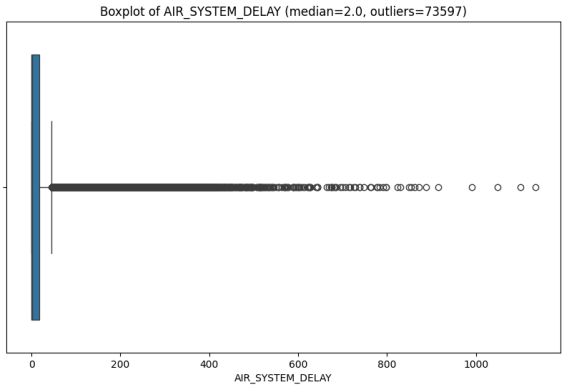

Boxplot of SCHEDULED_ARRIVAL (median=1520.0, outliers=0)
0
500
1000
1500
2000
2500
SCHEDULED_ARRIVAL

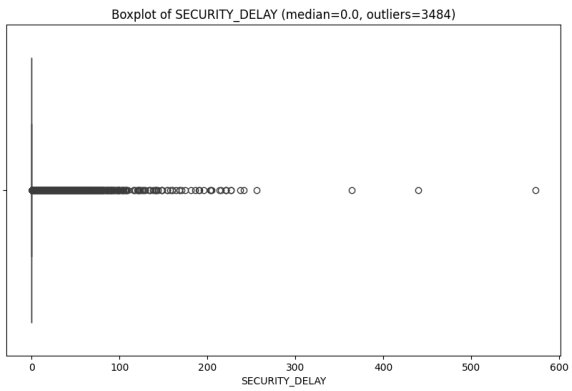

Boxplot of ARRIVAL_TIME (median=1512.0, outliers=0)
0
500
1000
1500
2000
2500
ARRIVAL_TIME

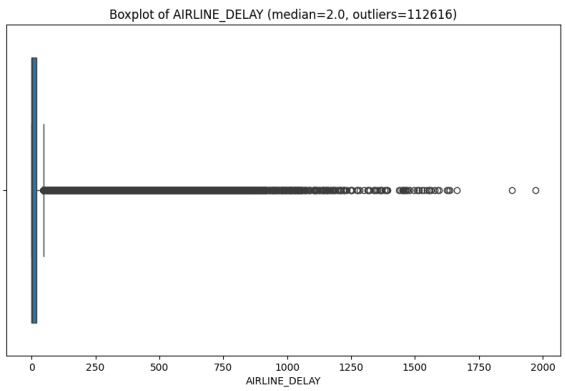


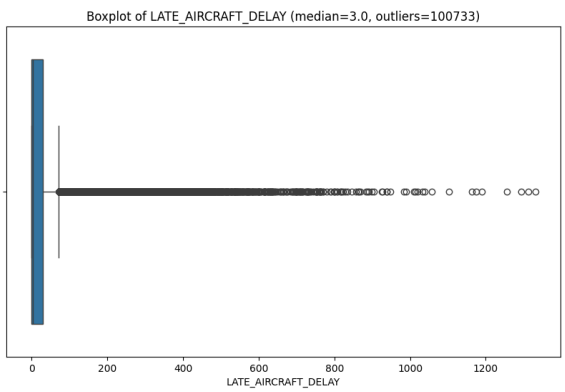


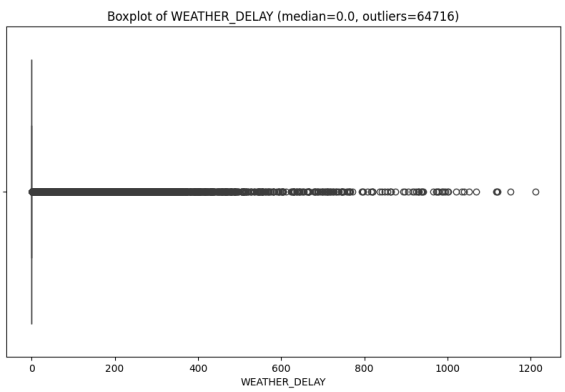












Nazwa	Mediana	Punkty oddalone
AIR_SYSTEM_DELAY	2	73 597
AIR_TIME	94	296 342

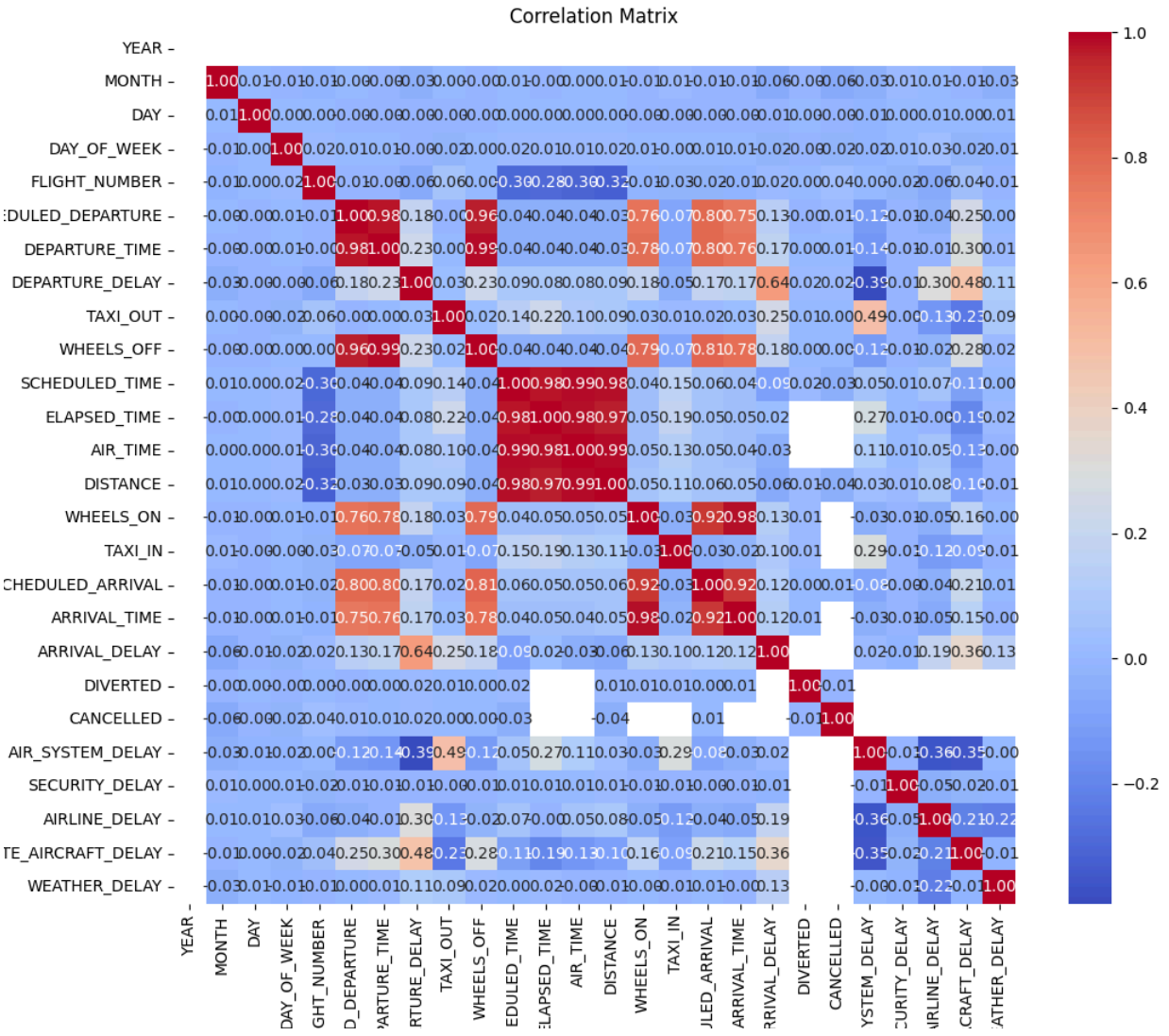
Nazwa	Mediana	Punkty oddalone
AIRLINE_DELAY	2	112 616
ARRIVAL_DELAY	−5	512 002
ARRIVAL_TIME	1 512	0
CANCELLED	0	89 884
DAY_OF_WEEK	4	0
DAY	16	0
DEPARTURE_DELAY	−2	736 242
DEPARTURE_TIME	1 330	0
DISTANCE	647	349 511
DIVERTED	0	15 187
ELAPSED_TIME	118	291 084
FLIGHT_NUMBER	1 690	27 073
LATE_AIRCRAFT_DELAY	3	100 733
MONTH	7	0
SCHEDULED_ARRIVAL	1 520	0
SCHEDULED_DEPARTURE	1 325	0
SCHEDULED_TIME	123	299 011
SECURITY_DELAY	0	3 484
TAXI_IN	6	282 538
TAXI_OUT	14	282 602
WEATHER_DELAY	0	64 716
WHEELS_OFF	1 343	0
WHEELS_ON	1 509	0
YEAR	2015	0

Na podstawie przedstawionych median oraz liczby punktów oddalonych dla poszczególnych atrybutów można sformułować następujące wnioski:

- Dla większości atrybutów numerycznych, takich jak AIR_SYSTEM_DELAY, AIRLINE_DELAY, DEPARTURE_DELAY czy LATE_AIRCRAFT_DELAY, mediana wynosi 0, 2 lub wartości bliskie zeru. Oznacza to, że typowy lot nie doświadcza istotnych opóźnień z tych przyczyn, a większość lotów przebiega zgodnie z planem.
- Wysoka liczba punktów oddalonych (np. 736 242 dla DEPARTURE_DELAY, 299 011 dla SCHEDULED_TIME, 296 342 dla AIR_TIME) wskazuje na obecność nietypowych, ekstremalnych przypadków w danych. Takie wartości mogą być wynikiem wyjątkowych zdarzeń, np. bardzo dużych opóźnień, awarii lub specyficznych tras.
- Mediana opóźnienia przylotu (ARRIVAL_DELAY) jest ujemna (-5), co sugeruje, że ponad połowa lotów przylatuje przed planowanym czasem lub z minimalnym opóźnieniem.
- Atrybuty binarne, takie jak CANCELLED czy DIVERTED, mają medianę 0, co oznacza, że większość lotów nie jest odwoływana ani przekierowywana.

- Dla atrybutów czasowych (np. ARRIVAL_TIME, DEPARTURE_TIME, WHEELS_ON, WHEELS_OFF) mediany odpowiadają typowym godzinom operacji lotniczych, a brak punktów oddalonych sugeruje, że wartości te są stabilne.
- Wysoka liczba punktów oddalonych w atrybutach związanych z czasem trwania lotu (AIR_TIME, ELAPSED_TIME, SCHEDULED_TIME) oraz dystansem (DISTANCE) odzwierciedla zróżnicowanie tras – od krótkich po bardzo długie loty.

5.3. Macierz korelacji



Istnieje silna korelacja pomiędzy atrybutami WHEELS_OFF i ARRIVAL_TIME (0.78), co pokazuje że większość opóźnij wynika z opóźnieniami na lądzie. Same przeloty są punktualne. Podobne wnioski można wysnuć w przypadku atrybutów SECURITY_DELAY i WEATHER_DELAY które prawie nie wykazują korelacji z całkowitym opóźnieniem co może sugerować że czasy odprawy i pogoda są niewielkim czynnikiem wpływającym na opóźnienia.

6. Podsumowanie

Na początku przeprowadzono analizę rozkładów atrybutów, z której wynika, że większość zmiennych numerycznych nie ma rozkładu normalnego — najczęściej przyjmują one postać rozkładów

prawoskośnych. Dodatkowo klasy zmiennej celu (czy lot jest opóźniony) są niezbalansowane — znacznie więcej lotów kończy się punktualnie lub z niewielkim opóźnieniem. W związku z nienormalnością rozkładów zastosowano współczynnik korelacji rang Spearmana do analizy zależności między zmiennymi.

Analiza korelacji wykazała istnienie silnych związków pomiędzy niektórymi grupami atrybutów. Szczególnie silna korelacja występuje między momentem startu (WHEELS_OFF) a czasem przylotu (ARRIVAL_TIME), co sugeruje, że opóźnienia przylotu wynikają przede wszystkim z opóźnień przy starcie, a sam czas przelotu jest relatywnie stabilny. Z kolei atrybuty takie jak WEATHER_DELAY czy SECURITY_DELAY wykazują bardzo niską korelację z opóźnieniem przylotu, co może świadczyć o ich mniejszym znaczeniu w kontekście predykcji.

Oceniono również jakość danych. Występuje bardzo duża liczba punktów odstających w wielu atrybutach (np. DEPARTURE_DELAY, AIR_TIME, SCHEDULED_TIME), co może wskazywać na nietypowe sytuacje operacyjne, takie jak intensywne warunki pogodowe, awarie techniczne czy długodystansowe trasy. Mimo to, z racji potencjalnej informacyjności takich obserwacji, zdecydowano się ich nie usuwać, ponieważ mogą mieć istotne znaczenie dla budowy modelu predykcyjnego.

Na podstawie powyższej analizy można stwierdzić, że dane są wystarczająco dobrej jakości, by możliwe było osiągnięcie postawionego celu eksploracji – predykcji opóźnień przylotów z dokładnością co najmniej 85%. Szczególna uwaga zostanie położona na maksymalizację czułości modelu, aby ograniczyć liczbę przypadków, w których rzeczywiste opóźnienie nie zostanie wykryte.