Eksploracja danych - etap 1

2015 Flight Delays and Cancellations

30.05.2025

Krzysztof Nasuta 193328, Filip Dawidowski 193433, Aleks Iwicki 193354

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**.

Ogólny opis zbioru (ii)

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.

```
YEAR,MONTH,DAY,DAY_OF_WEEK,AIRLINE,FLIGHT_NUMBER,TAIL_NUMBER,ORIGIN_AIRPORT,DESTINATION...
2015,1,1,4,AS,98,N407AS,ANC,SEA...
2015,1,1,4,AA,2336,N3KUAA,LAX,PBI...
2015,1,1,4,US,840,N171US,SFO,CLT...
```

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%**.

Określenie celu eksploracji i kryteriów sukcesu (ii)

W naszym przypadku chcemy skupić się na **maksymalizacji czułości modelu ponad swoistość**, ponieważ błędne wykrycie opóźnienia nie jest tak istotne, jak przeoczenie rzeczywistego opóźnienia.

Charakterystyka zbioru

- Pochodzenie: Kaggle
- Liczba przykładów: **5 819 080**
- Format: CSV (3 pliki: `flights.csv` - właściwy zbiór, `airports.csv` - informacje o lotniskach, `airlines.csv` - informacje o liniach lotniczych)
- Ilość zbiorów danych: 1

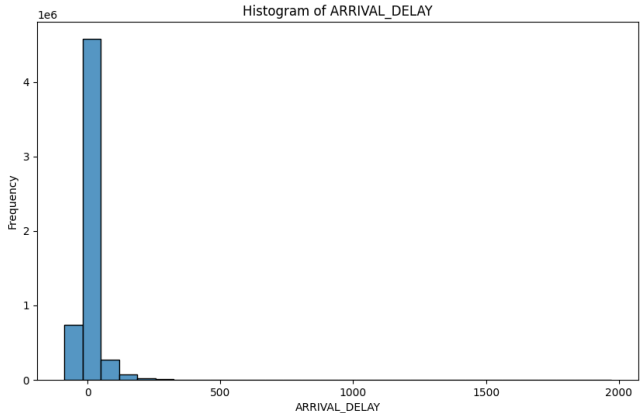
Brakujące dane

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.

W zbiorze **nie** wystąpiły dane niezrozumiałe.

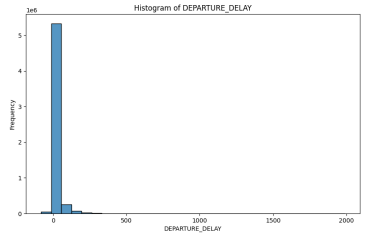
Wynik eksploracyjnej analizy danych

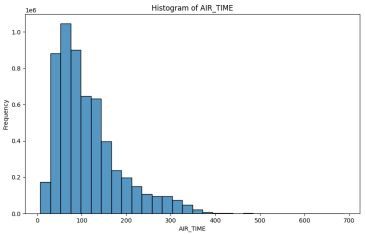
Rozkłady wartości atrybutów

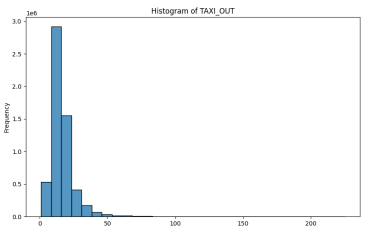


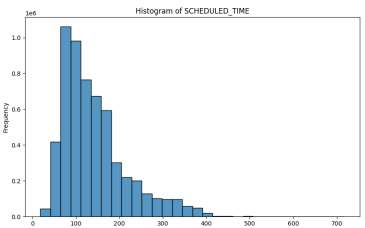
W przypadku atrybutu ARRIVAL_DELAY zauważalna jest **przewaga lotów punktualnych** lub z niewielkim opóźnieniem (poniżej 15 minut) względem lotów znacząco opóźnionych. Klasy są **niezbalansowane**, co może wpłynąć na skuteczność modeli klasyfikacyjnych.

Rozkłady wartości atrybutów (ii)



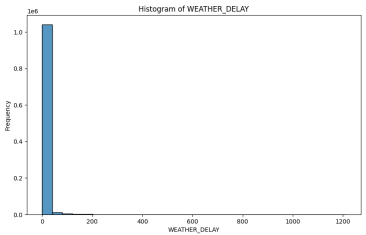


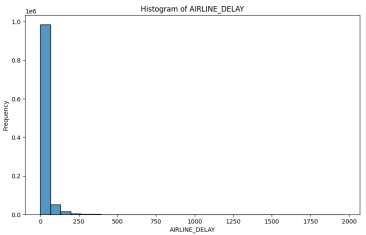


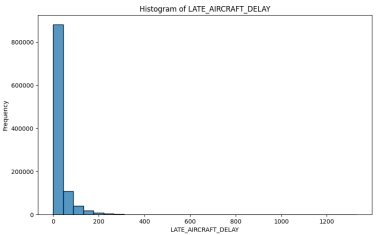


Rozkłady większości atrybutów numerycznych, takich jak DEPARTURE_DELAY, AIR_TIME, TAXI_OUT czy SCHEDULED_TIME, nie przypominają rozkładu normalnego. Najczęściej obserwujemy **rozkład prawoskośny** - większość wartości skupia się w niższych przedziałach, a ogon rozkładu jest wydłużony w stronę wyższych wartości.

Rozkłady wartości atrybutów (iii)

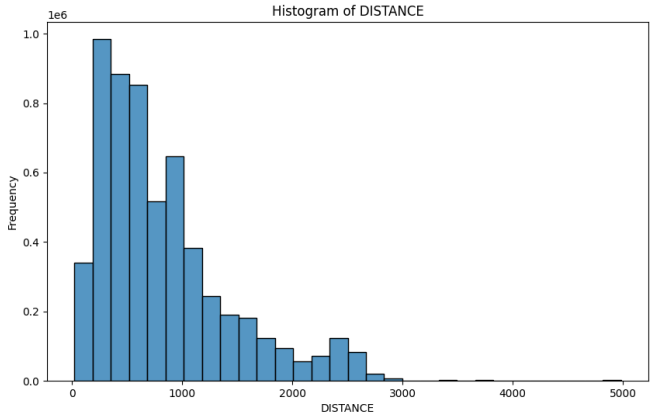






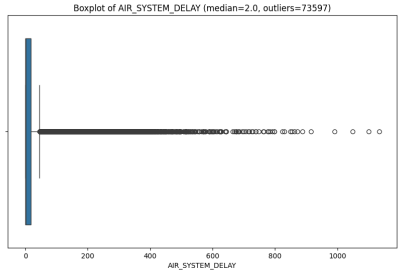
Wysokie, rzadko występujące wartości w atrybutach takich jak WEATHER_DELAY, AIRLINE_DELAY oraz LATE_AIRCRAFT_DELAY mogą wskazywać na **zdarzenia nietypowe**, takie jak intensywne burze, problemy techniczne lub opóźnienia łańcuchowe wynikające z wcześniejszych lotów. Tego typu przypadki mają istotne znaczenie dla analizy przyczyn opóźnień i mogą być kluczowe przy budowie predykcyjnych modeli.

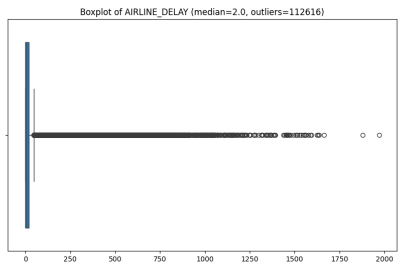
Rozkłady wartości atrybutów (iv)



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.**

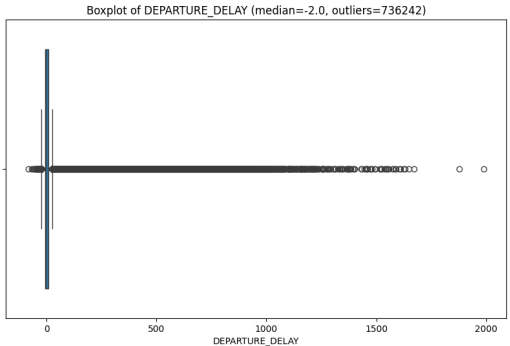
Punkty oddalone

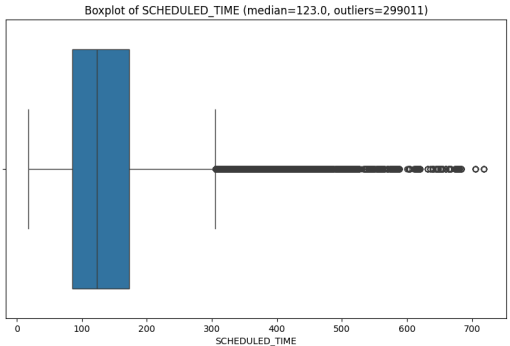




Dla większości atrybutów numerycznych, takich jak AIR_SYSTEM_DELAY, AIRLINE_DELAY, DEPARTURE_DELAY czy LATE_AIRCRAFT_DELAY, mediana wynosi 0, 2 lub wartości bliskie zeru. Oznacza to, że **typowy lot nie doświadcza istotnych opóźnień** z tych przyczyn, a większość lotów przebiega zgodnie z planem.

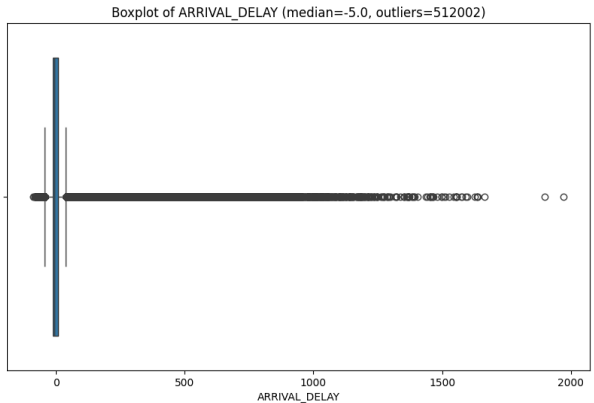
Punkty oddalone (ii)





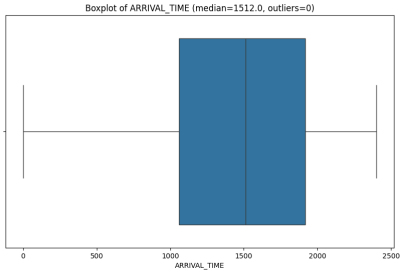
Wysoka liczba punktów oddalonych (np. 736 242 dla DEPARTURE_DELAY, 299 011 dla SCHEDULED_TIME, 296 342 dla AIR_TIME) wskazuje na obecność **nietypowych, ekstremalnych przypadków w danych**. Takie wartości mogą być wynikiem **wyjątkowych** zdarzeń, np. bardzo dużych opóźnień, awarii lub specyficznych tras.

Punkty oddalone (iii)

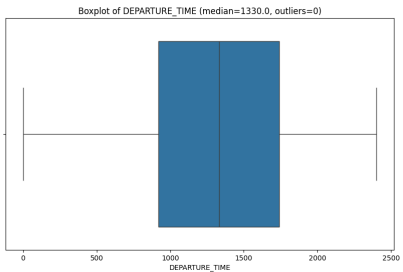


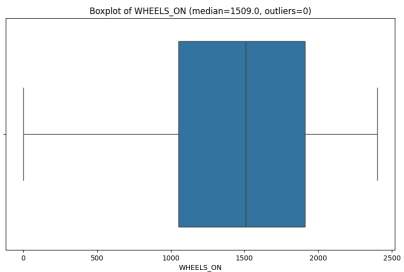
Mediana opóźnienia przylotu (ARRIVAL_DELAY) jest ujemna (-5), co sugeruje, że ponad połowa lotów przylatuje przed planowanym czasem lub z minimalnym opóźnieniem.

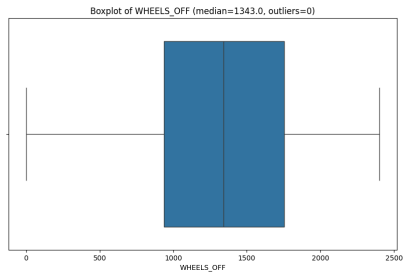
Punkty oddalone (iv)



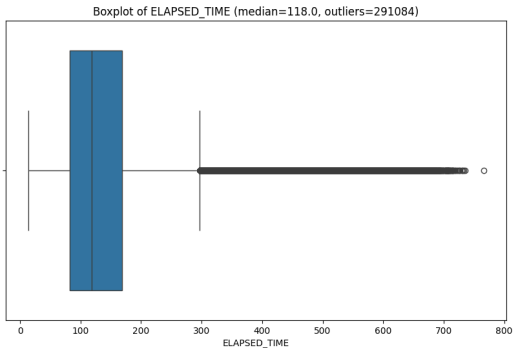
Dla atrybutów czasowych (np. ARRIVAL_TIME, DEPARTURE_TIME, WHEELS_ON, WHEELS_OFF) mediany odpowiadają typowym godzinom operacji lotniczych, a **brak punktów oddalonych** sugeruje, że wartości te są **stabilne**.

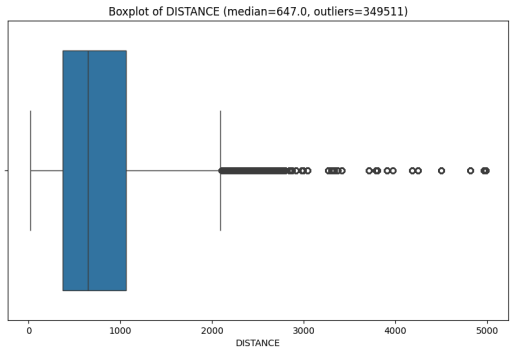






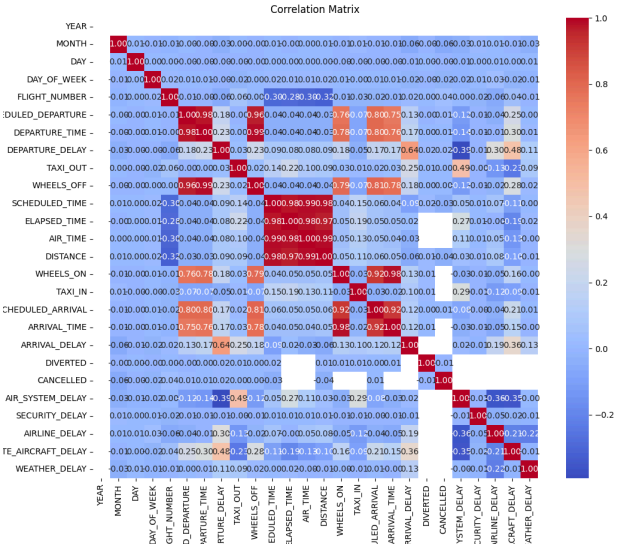
Punkty oddalone (v)





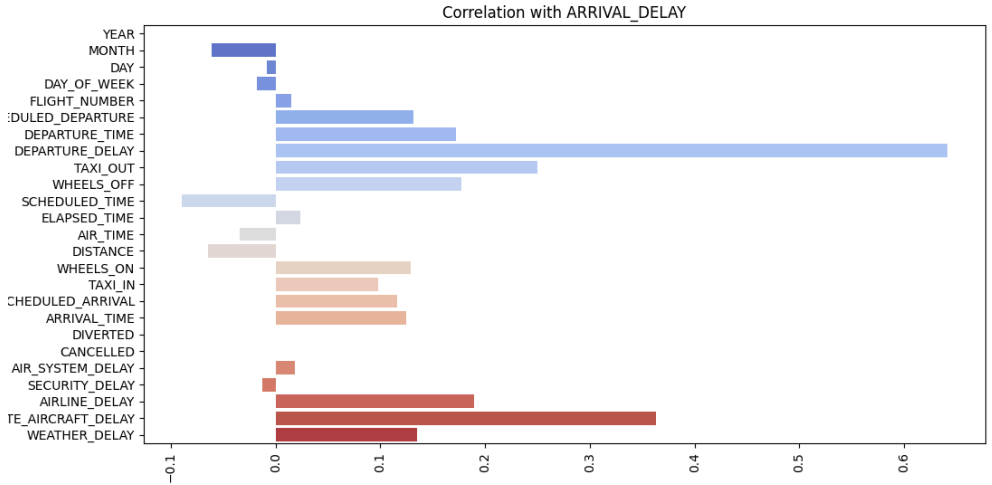
Wysoka liczba punktów oddalonych w atrybutach związanych z czasem trwania lotu (AIR_TIME, ELAPSED_TIME, SCHEDULED_TIME) oraz dystansem (DISTANCE) odzwierciedla zróżnicowanie tras - od krótkich po bardzo długie loty.

Korelacje



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.

Korelacje (ii)



Korelacja innych pól z atrybutem celu ARRIVAL_DELAY.

Wnioski

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.

Podsumowanie (ii)

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.

Podsumowanie (iii)

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.

Podsumowanie (iv)

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.

Dziękujemy za uwagę!