ANALIZA DANYCH LOTNICZYCH

Prezentacja finalna

AGENDA

01 NASZ ZESPÓŁ

02 CELE I MOTYWACJE

03 ZBIORY DANYCH

04 METODY ANALIZY DANYCH

05 TECHNOLOGIE

06 BADANIA BIZNESOWE

07 BADANIA NAUKOWE

08 WNIOSKI

NASZ ZESPÓŁ

PAWEŁ GALEWICZ

BARTOSZ JURCZEWSKI

MATEUSZ MACIASZEK

PIOTR WARDĘCKI

CELE I MOTYWACJE

Chęć zbadania branży która jest integralną częścią współczesnego świata

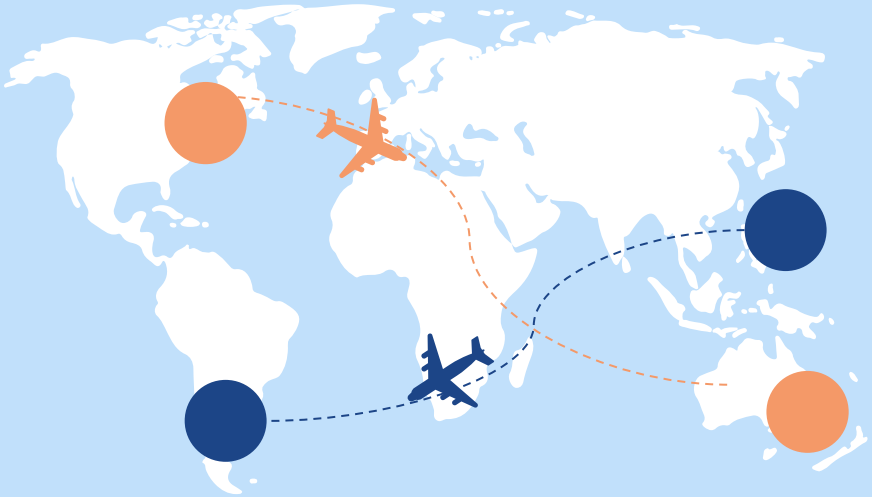
Sprawdzenie hipotezy, że największa częstotliwość lotów występuje w okresie wakacyjnym oraz przed świątecznym

Sprawdzenie wpływu epidemii covid-19 na częstotliwość lotów

Sprawdzenie tempa rozwoju branży na przestrzeni lat

Sprawdzenie średniego dystansu lotów

ZBIORY DANYCH



1. **Zbiór danych lotniczych**
- Departament transportu USA
- Lata 2009-2020
- 28 kolumn
- 9,5 GB danych
2. **Zbiór danych o zachorowaniach na COVID-19**
- Dane agregowane z oficjalnych danych stanowych
- 3 kolumny: liczba przypadków, stan, data
3. **Zbiór świąt federalnych 2009-2020**
- Rozszerzony przez nas o dni przed i po świętach
4. **Zbiór dat historycznych wydarzeń w USA 2009-2020**
- Zbudowany na podstawie listy najistotniejszych wydarzeń w dziejach USA

METODY ANALIZY DANYCH

Loty na przestrzeni dekady

- Statystyki długości dystansów i częstości lotów na przestrzeni lat
- Korelacja okresu roku i średniej długości lotu albo opóźnienia
- Częstotliwość lotów w okresach roku
- Zmiany statystyk w okolicach istotnych historycznie wydarzeń w USA
- Klastrowanie pod kątem opóźnień

Wpływ pandemii na loty

- Korelacja statystyk
- Zmiany w częstotliwości lotów ostatnim roku
- Model regresji z wykorzystaniem danych o zachorowaniach

STOS TECHNOLOGICZNY

PYTHON

SPARK

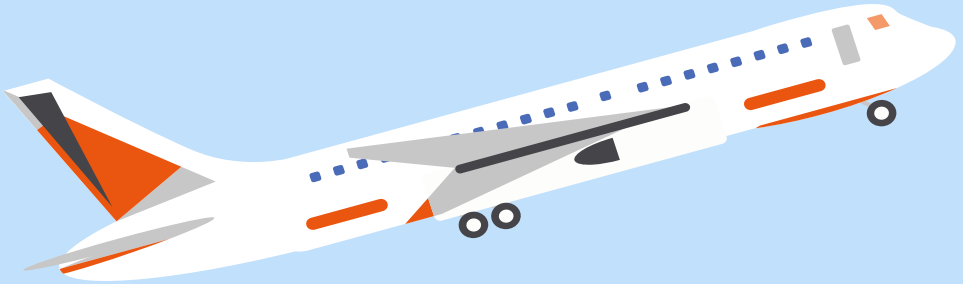
ANACONDA

DASK

JUPYTER

DOCKER

BADANIA BIZNESOWE



Statystyka Dystansu przebytego przez samolot

FL_DATE	OP_CARRIER	DEP_DELAY	ARR_DELAY	CANCELLED	DISTANCE
2019-04-04	NaN	-5.0	-9.0	0.0	5095.0
2019-04-05	NaN	-2.0	12.0	0.0	5095.0
2019-04-05	NaN	19.0	-2.0	0.0	5095.0
2019-04-06	NaN	-5.0	26.0	0.0	5095.0
2019-04-06	NaN	-3.0	-43.0	0.0	5095.0
2019-04-07	NaN	2.0	35.0	0.0	5095.0
2019-04-07	NaN	5.0	-10.0	0.0	5095.0
2019-04-08	NaN	13.0	48.0	0.0	5095.0
2019-04-08	NaN	3.0	-41.0	0.0	5095.0
2019-04-09	NaN	5.0	3.0	0.0	5095.0
2019-04-11	NaN	28.0	50.0	0.0	5095.0
2019-04-12	NaN	37.0	6.0	0.0	5095.0
2019-04-12	NaN	5.0	12.0	0.0	5095.0
2019-04-13	NaN	-1.0	-13.0	0.0	5095.0
2019-04-13	NaN	0.0	-13.0	0.0	5095.0
2019-04-14	NaN	-1.0	-30.0	0.0	5095.0
2019-04-14	NaN	2.0	30.0	0.0	5095.0
2019-04-15	NaN	2.0	-20.0	0.0	5095.0
2019-04-15	NaN	47.0	35.0	0.0	5095.0
2019-04-16	NaN	5.0	23.0	0.0	5095.0

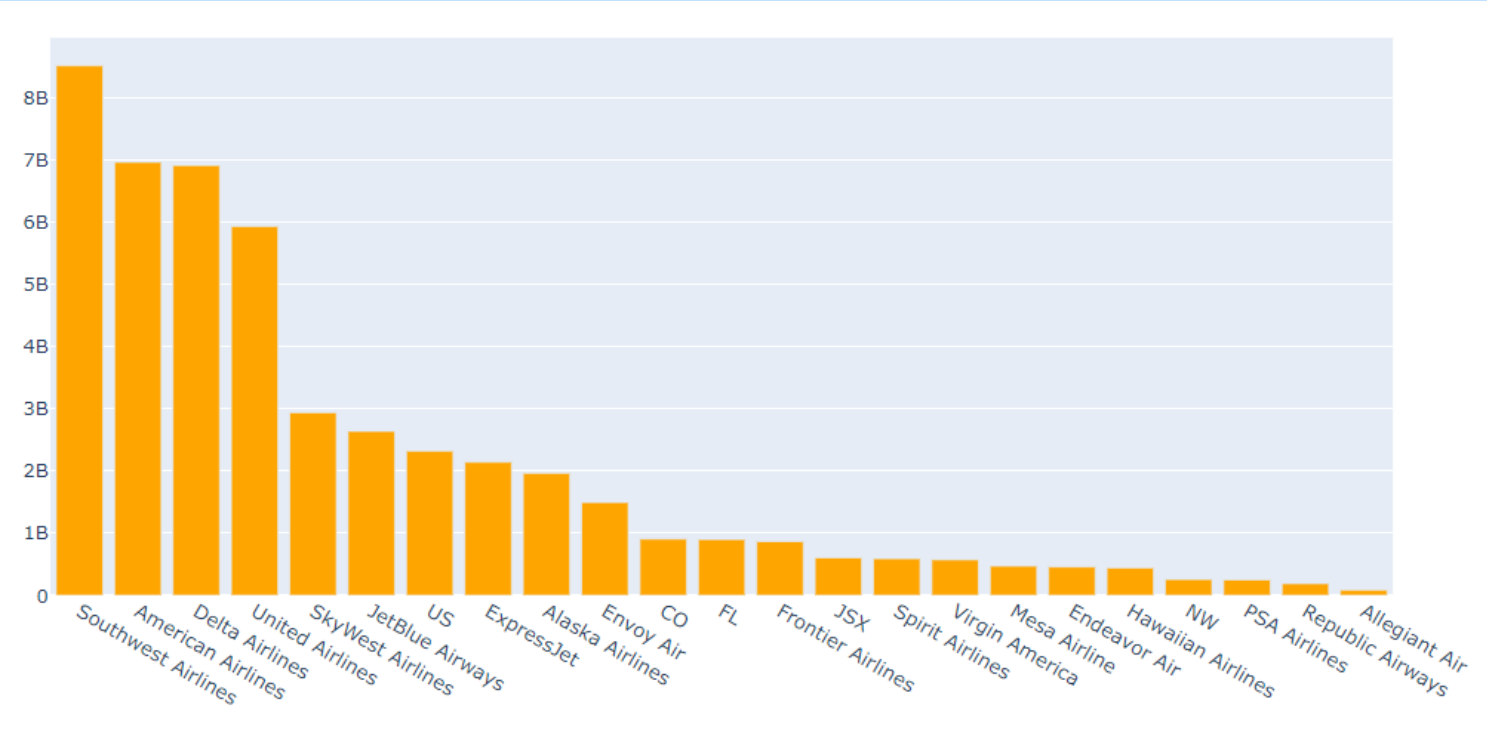
Średni dystans:

787

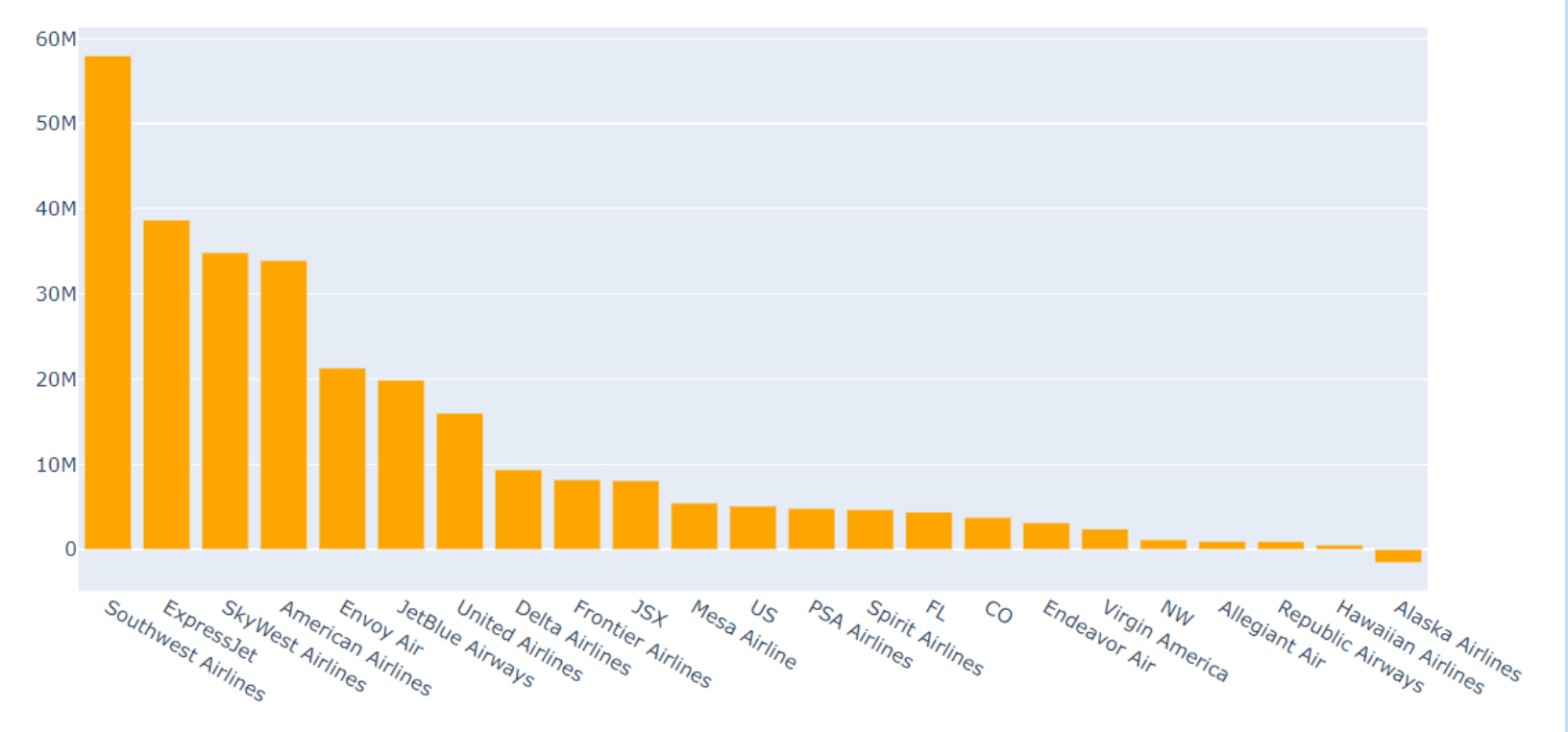
mile

FL_DATE	OP_CARRIER	DEP_DELAY	ARR_DELAY	CANCELLED	DISTANCE
2009-02-03	American Airlines	71.0	NaN	1.0	11.0
2009-07-07	PSA Airlines	65.0	35.0	0.0	11.0
2013-07-27	US	NaN	NaN	1.0	17.0
2009-01-21	American Airlines	-5.0	-5.0	0.0	21.0
2015-04-20	American Airlines	112.0	NaN	1.0	21.0
2012-12-10	American Airlines	1170.0	1182.0	0.0	24.0
2014-03-12	ExpressJet	NaN	NaN	1.0	24.0
2009-09-26	SkyWest Airlines	NaN	NaN	1.0	25.0
2009-09-26	SkyWest Airlines	NaN	NaN	1.0	25.0
2009-11-18	SkyWest Airlines	NaN	NaN	1.0	25.0
2016-11-28	SkyWest Airlines	232.0	NaN	0.0	25.0
2014-02-07	ExpressJet	38.0	NaN	0.0	26.0
2016-08-14	ExpressJet	7.0	56.0	0.0	28.0
2020-05-08	NaN	22.0	35.0	0.0	29.0
2020-05-09	NaN	0.0	12.0	0.0	29.0
2020-05-10	NaN	-3.0	7.0	0.0	29.0
2020-05-11	NaN	-9.0	NaN	0.0	29.0
2020-05-14	NaN	-1.0	1.0	0.0	29.0
2020-05-15	NaN	-3.0	8.0	0.0	29.0
2020-05-16	NaN	16.0	37.0	0.0	29.0

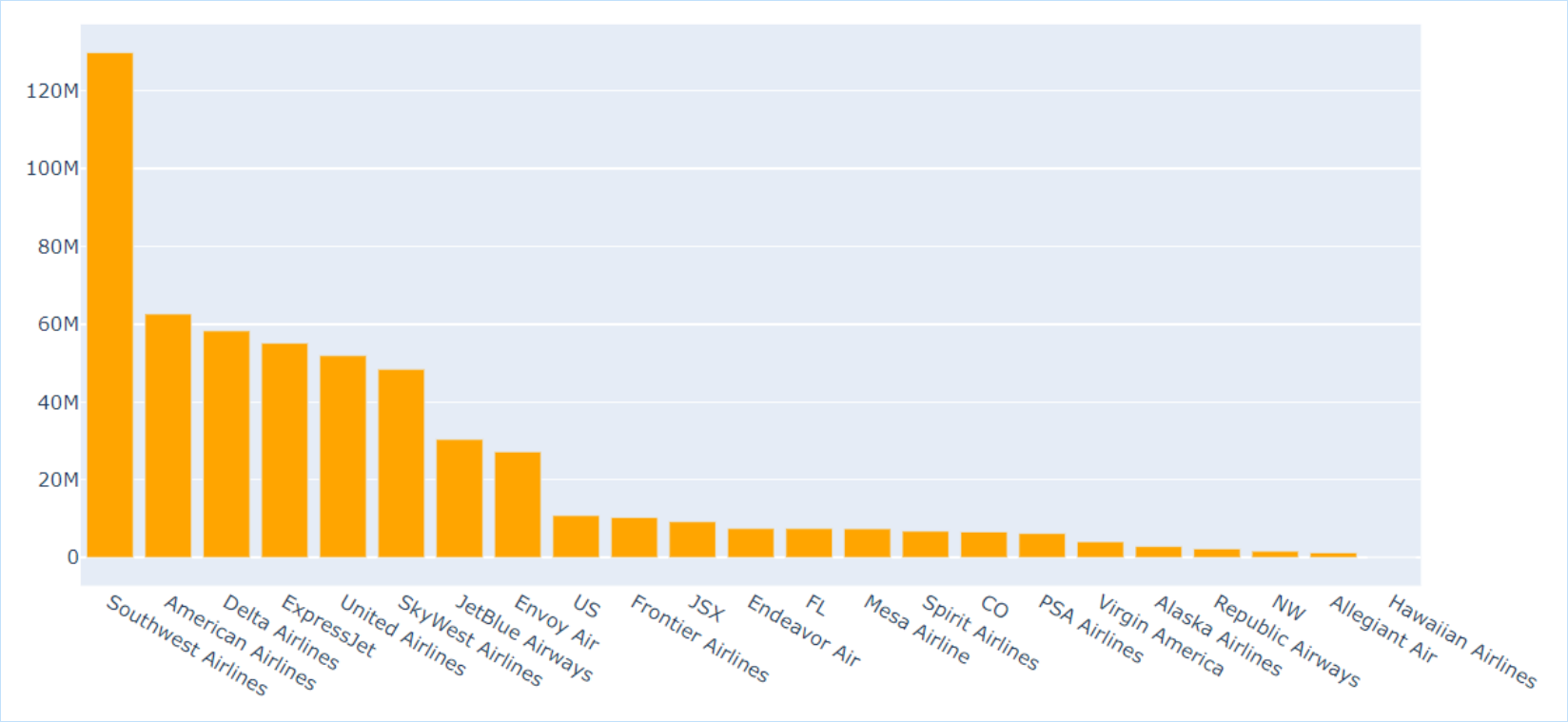
Całkowity dystans pokonany przez przewoźnika



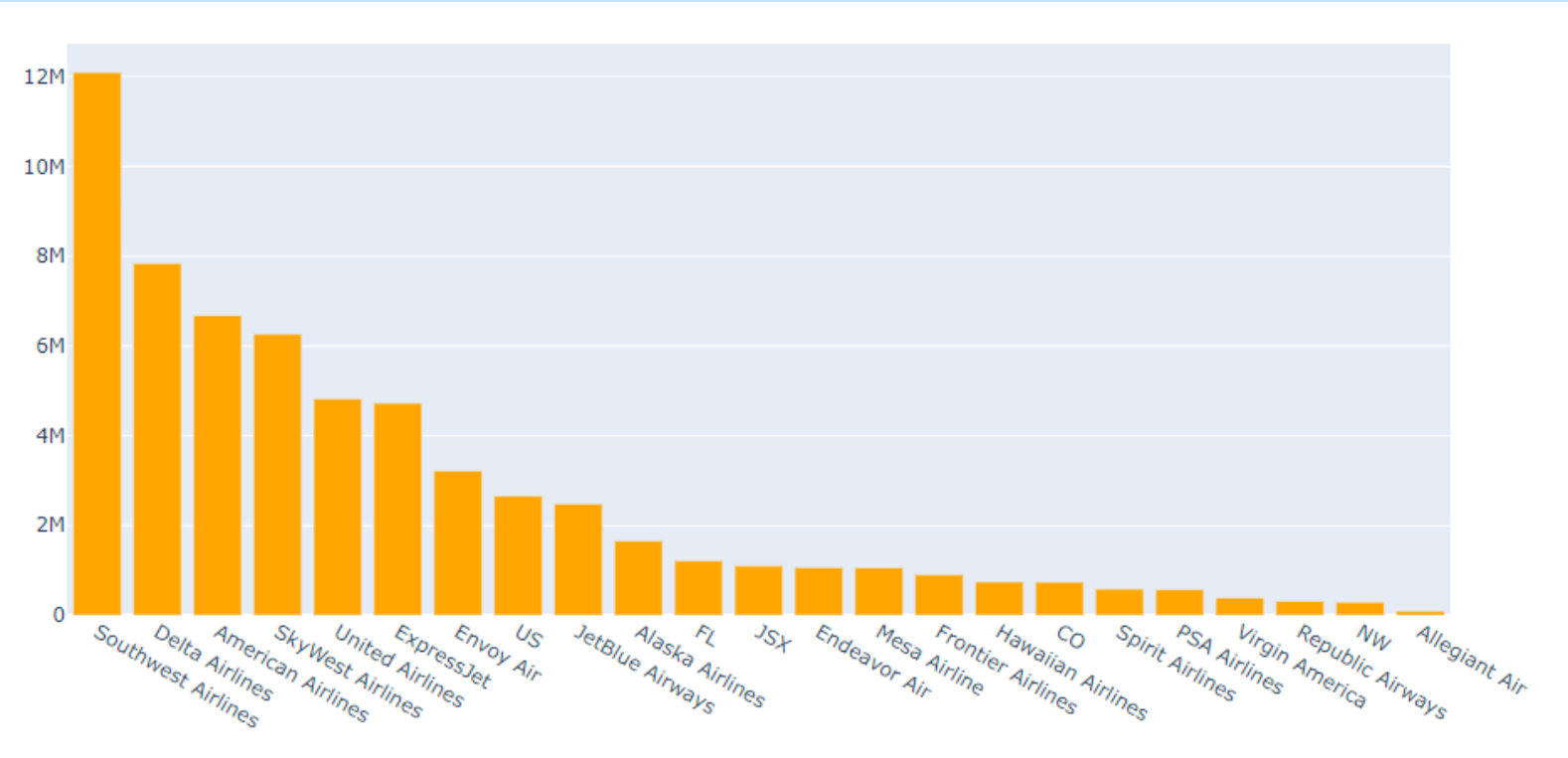
Opóźnienia przylotów przewoźnika



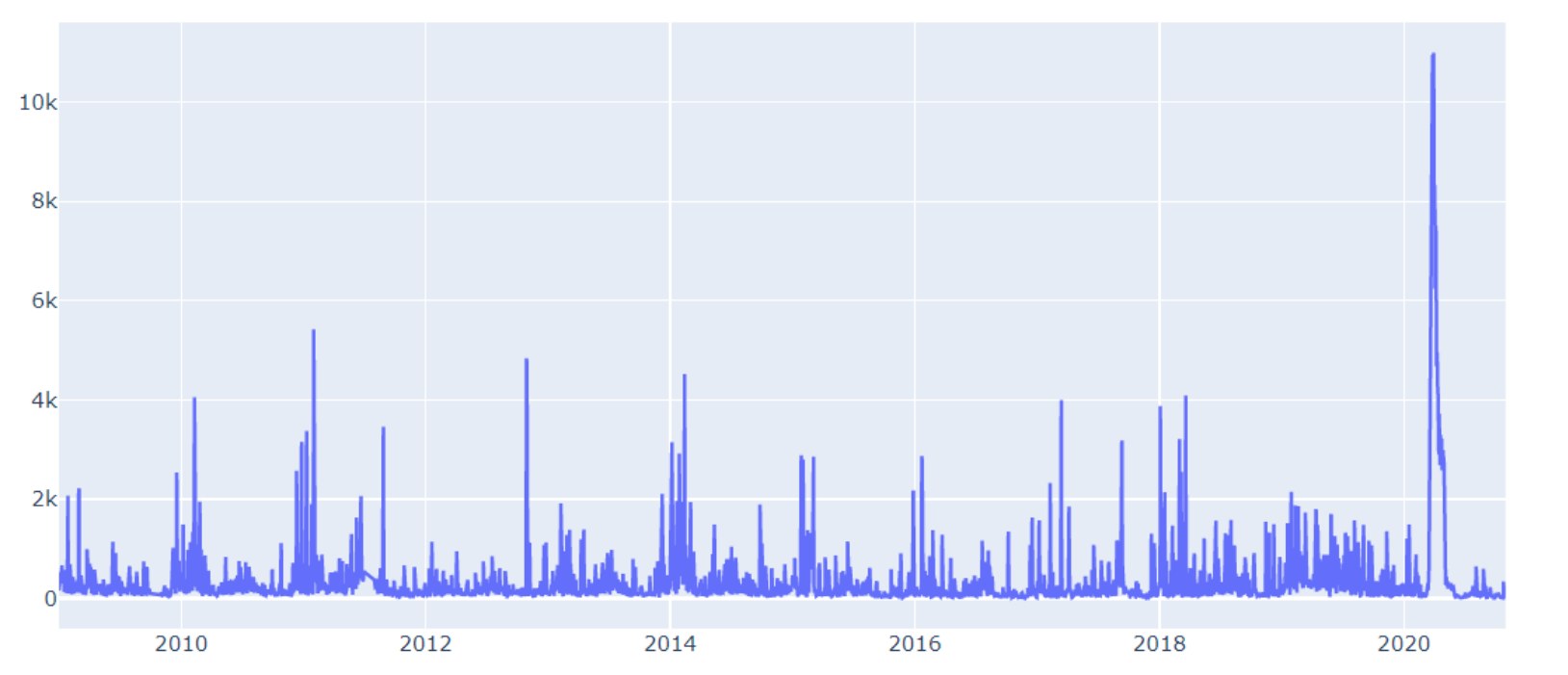
Opóźnienia odlotów przewoźnika



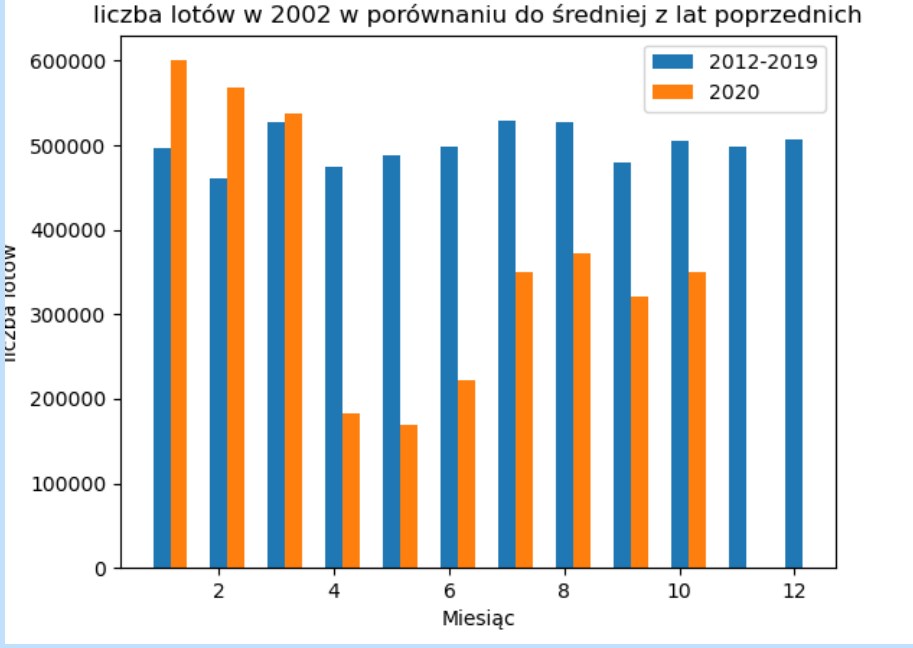
Najpopularniejszy przewoźnik

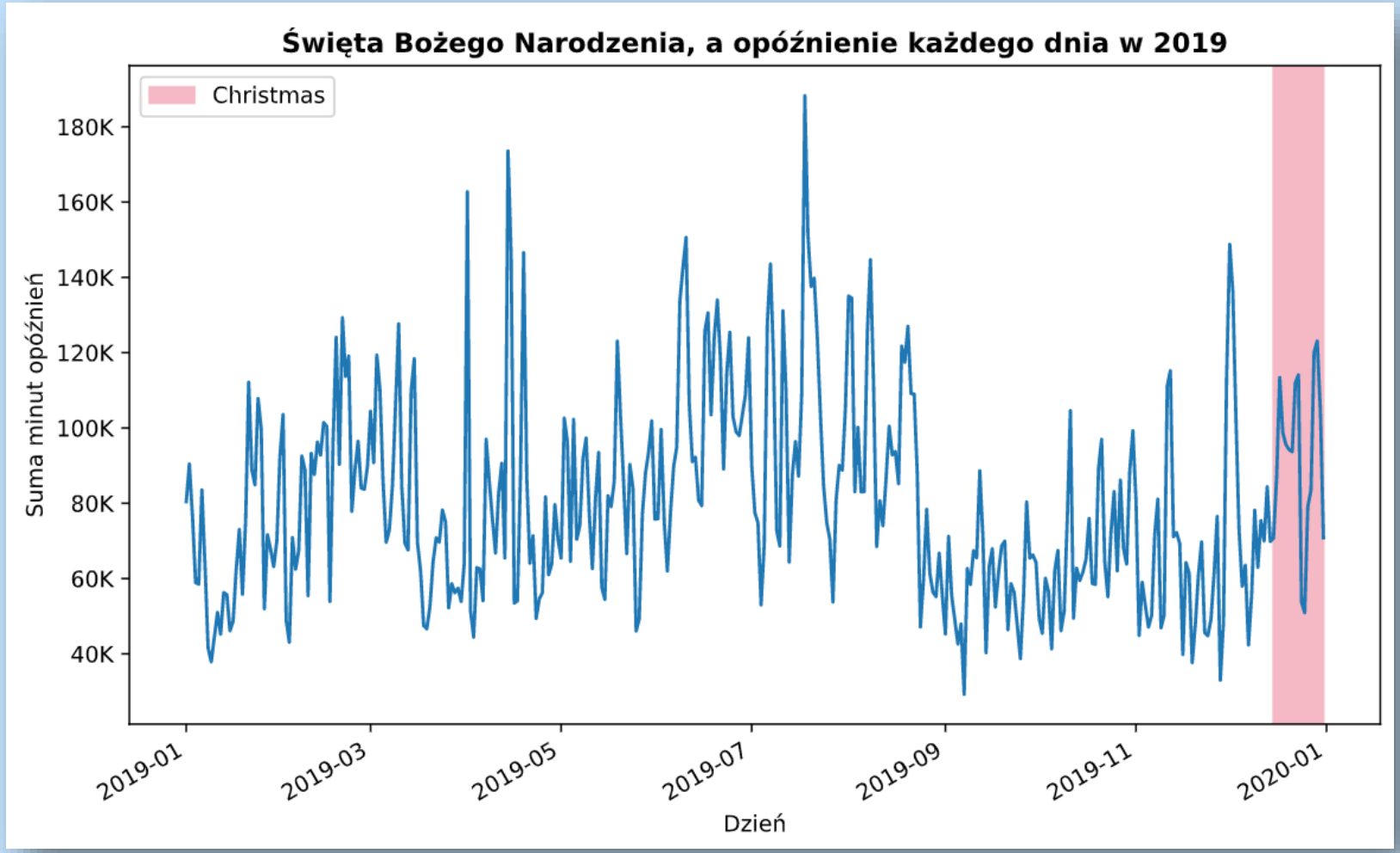


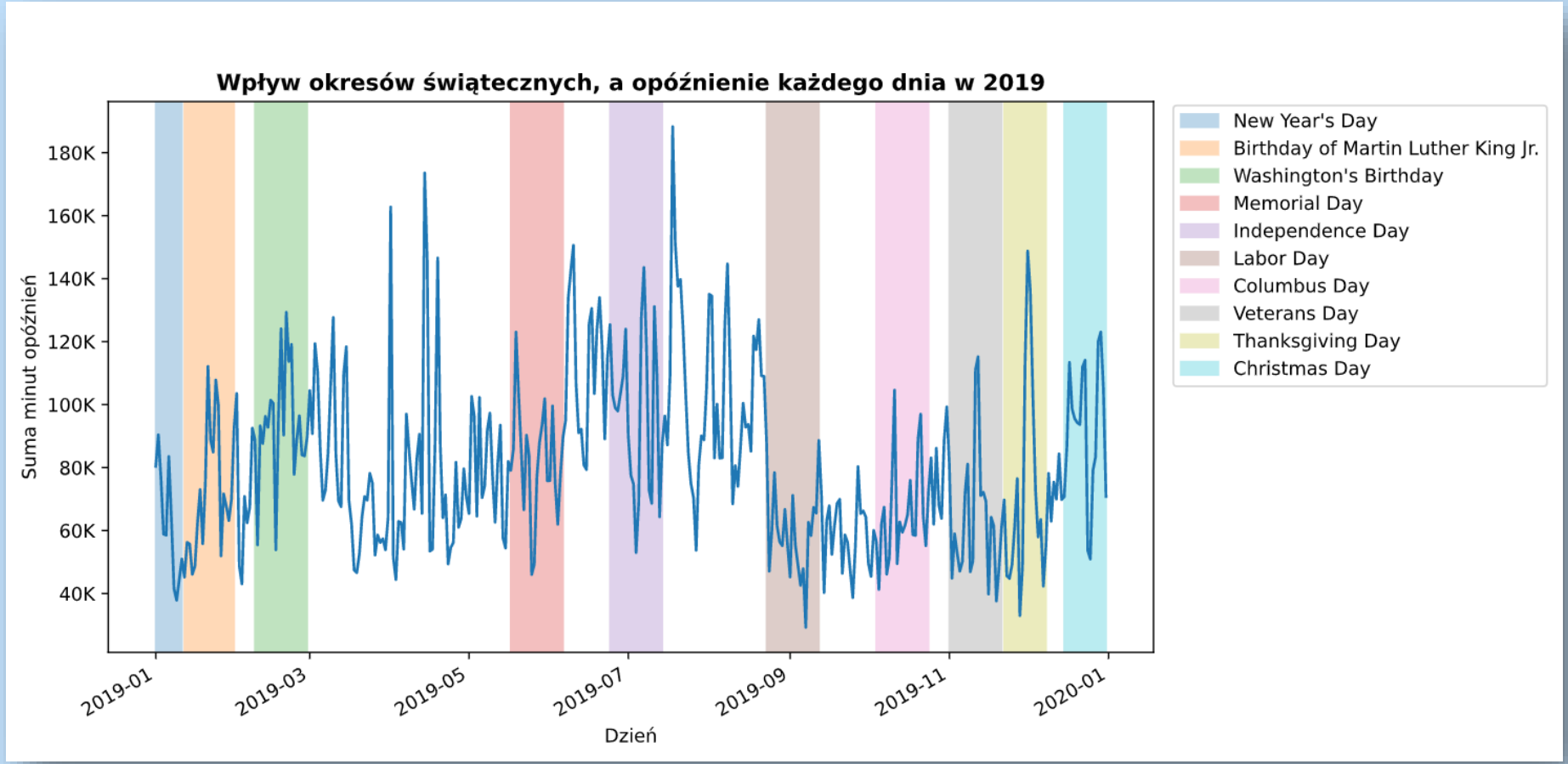
Odwołane loty 2009-2020

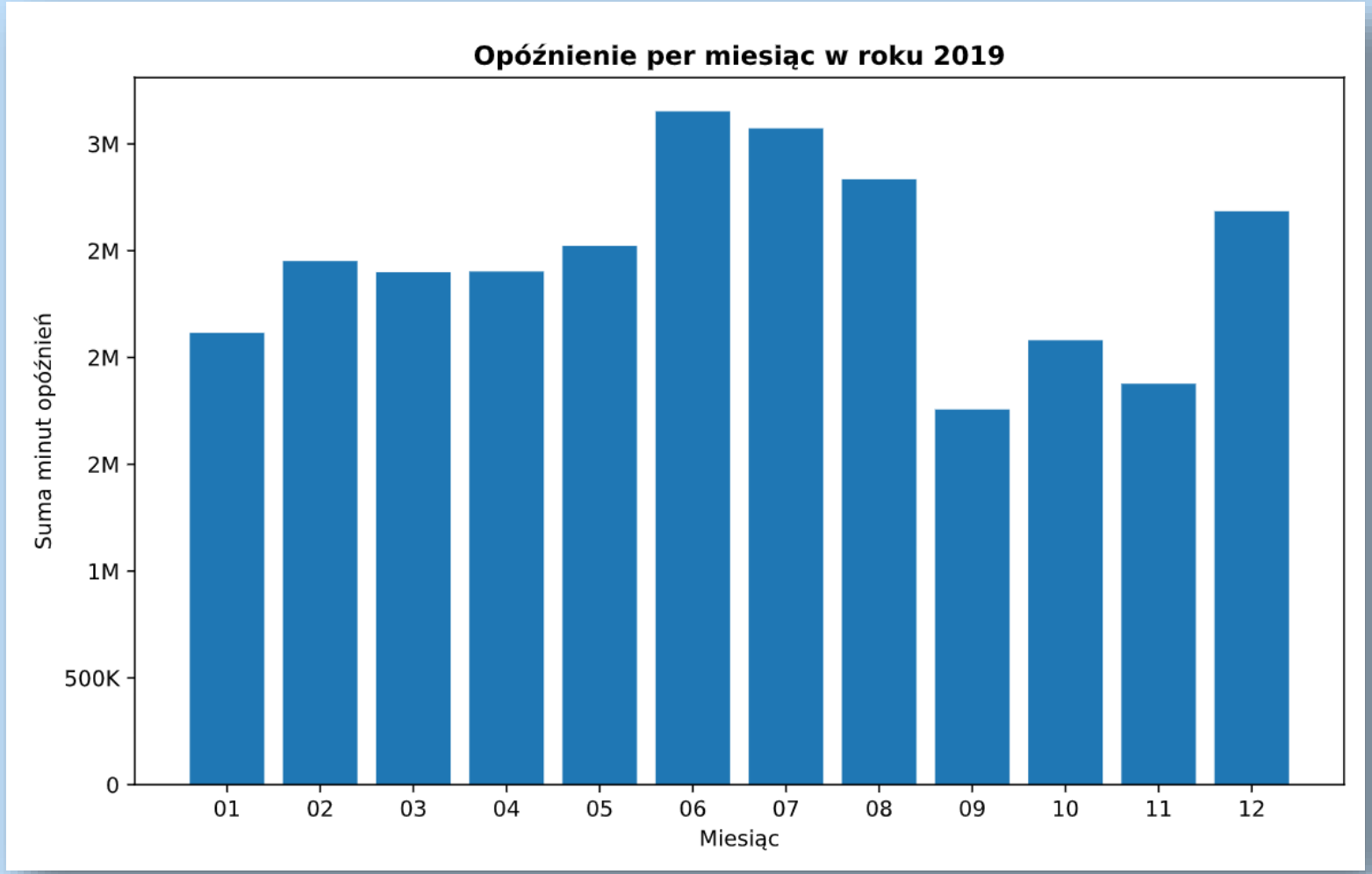


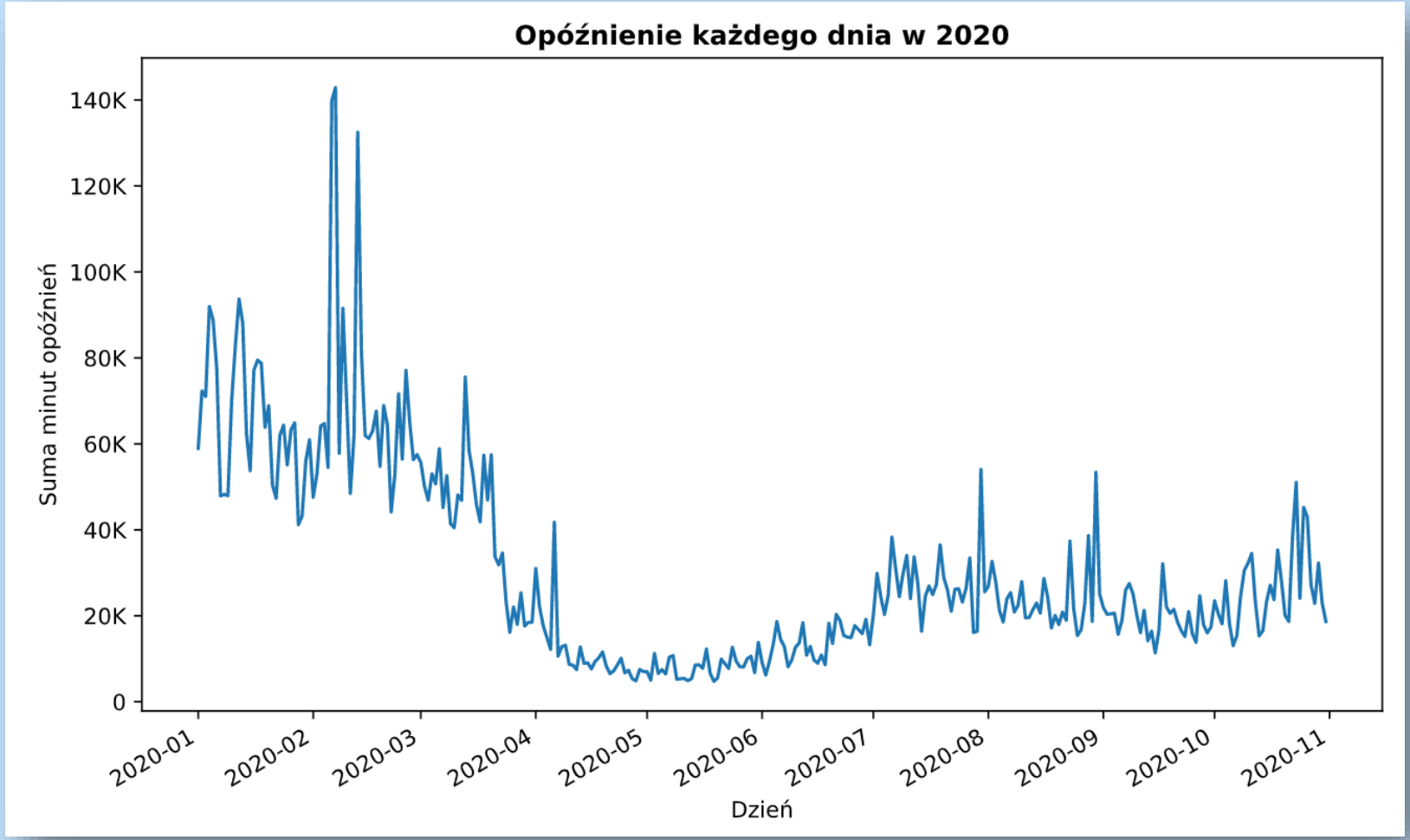
Wpływ COVID-19 na liczbę lotów

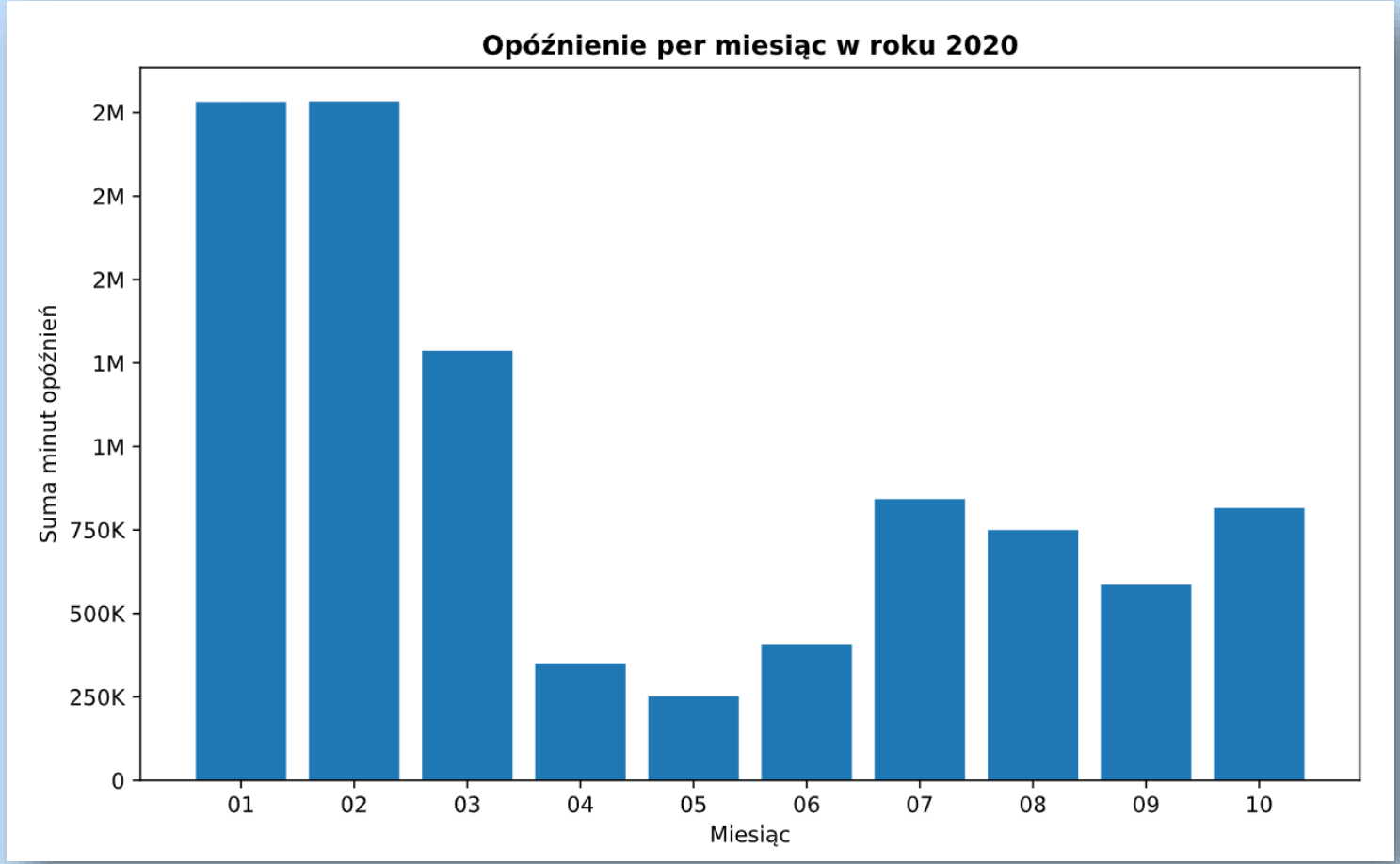


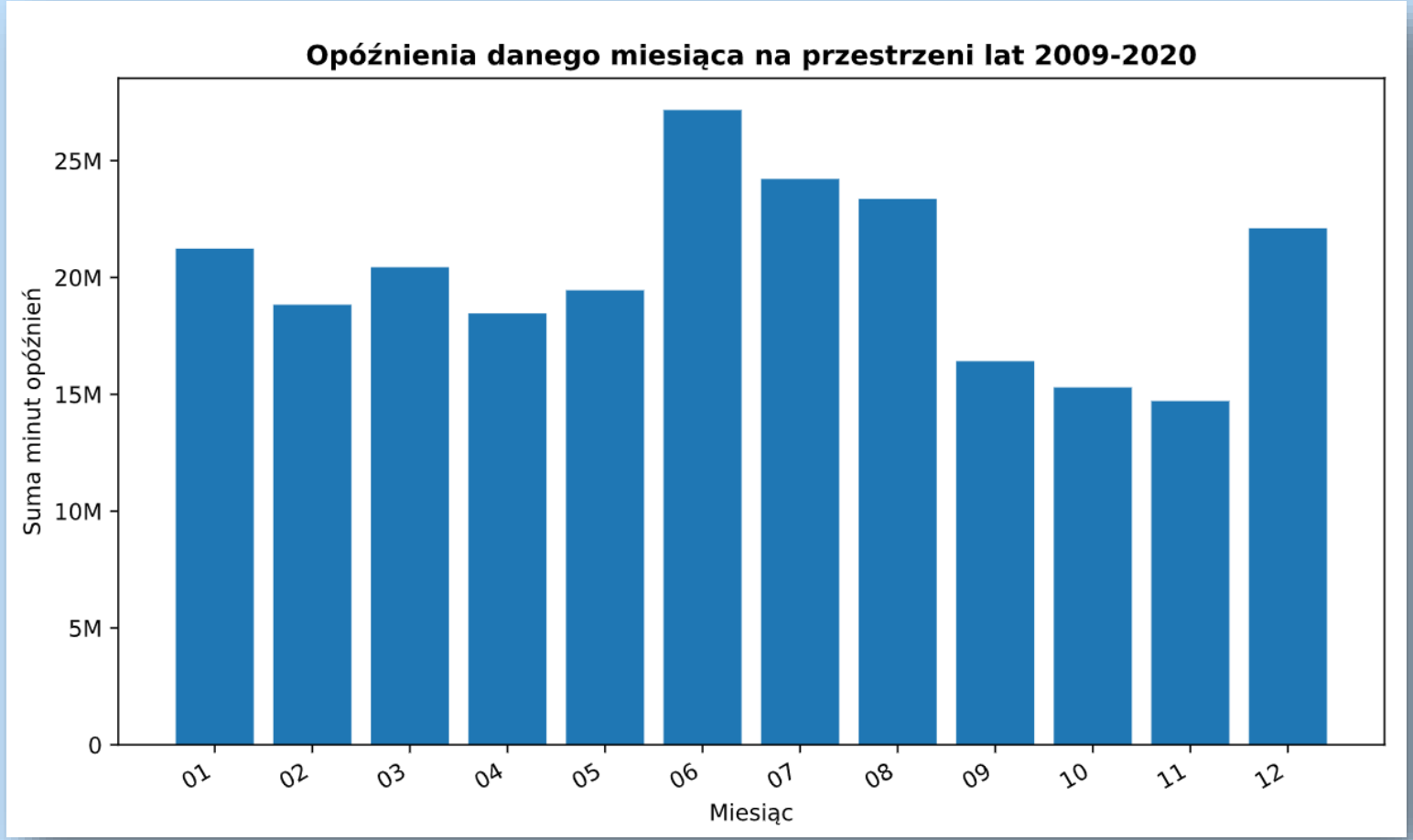

Święta Bożego Narodzenia, a opóźnienie każdego dnia w 2019
Christmas
180K
160K
140K
120K
100K
80K
60K
40K
Suma minut opóźnień
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
Dzień

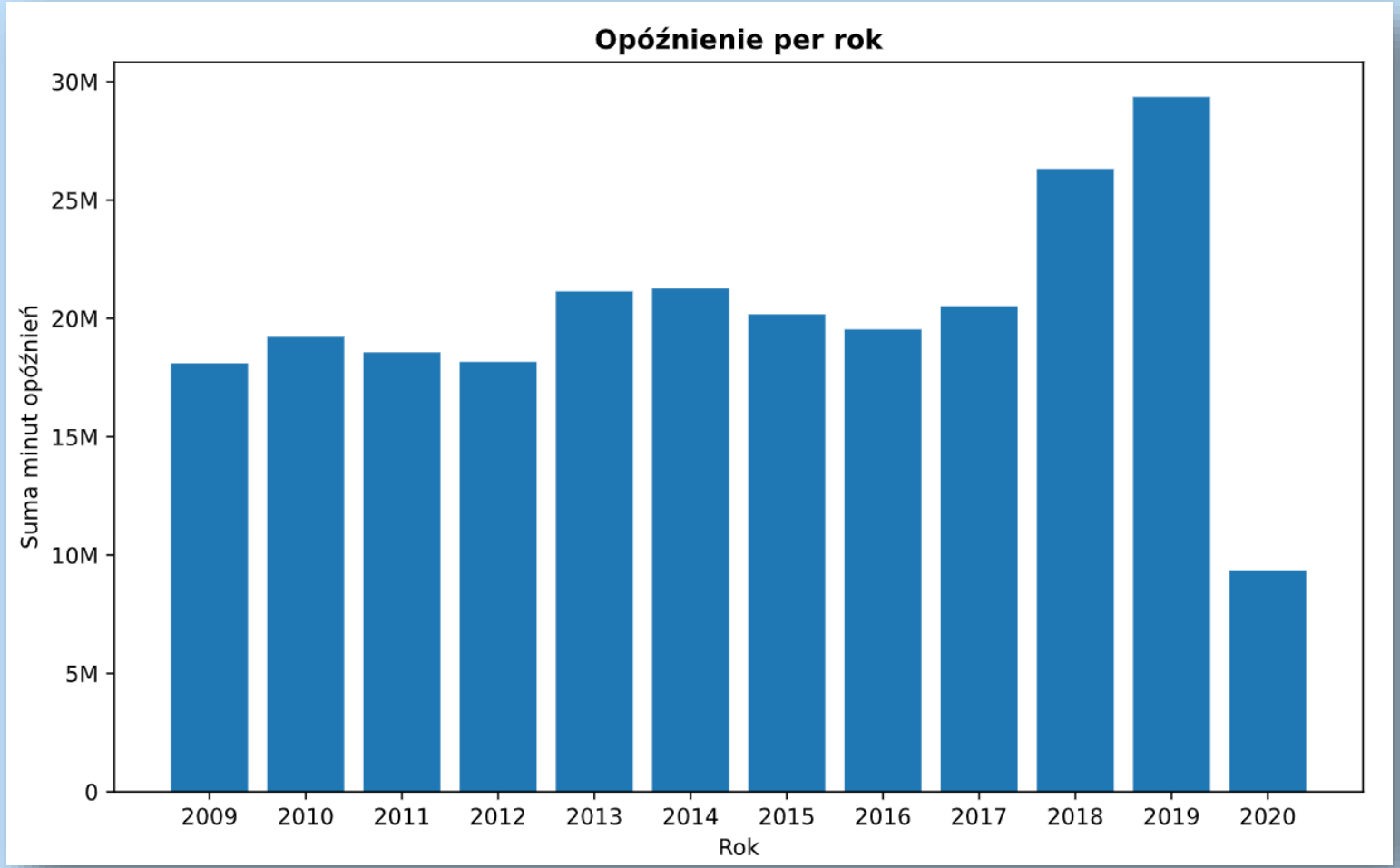

Wpływ okresów świątecznych, a opóźnienie każdego dnia w 2019
Suma minut opóźnień
180K
160K
140K
120K
100K
80K
60K
40K
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
Dzień
New Year's Day
Birthday of Martin Luther King Jr.
Washington's Birthday
Memorial Day
Independence Day
Labor Day
Columbus Day
Veterans Day
Thanksgiving Day
Christmas Day

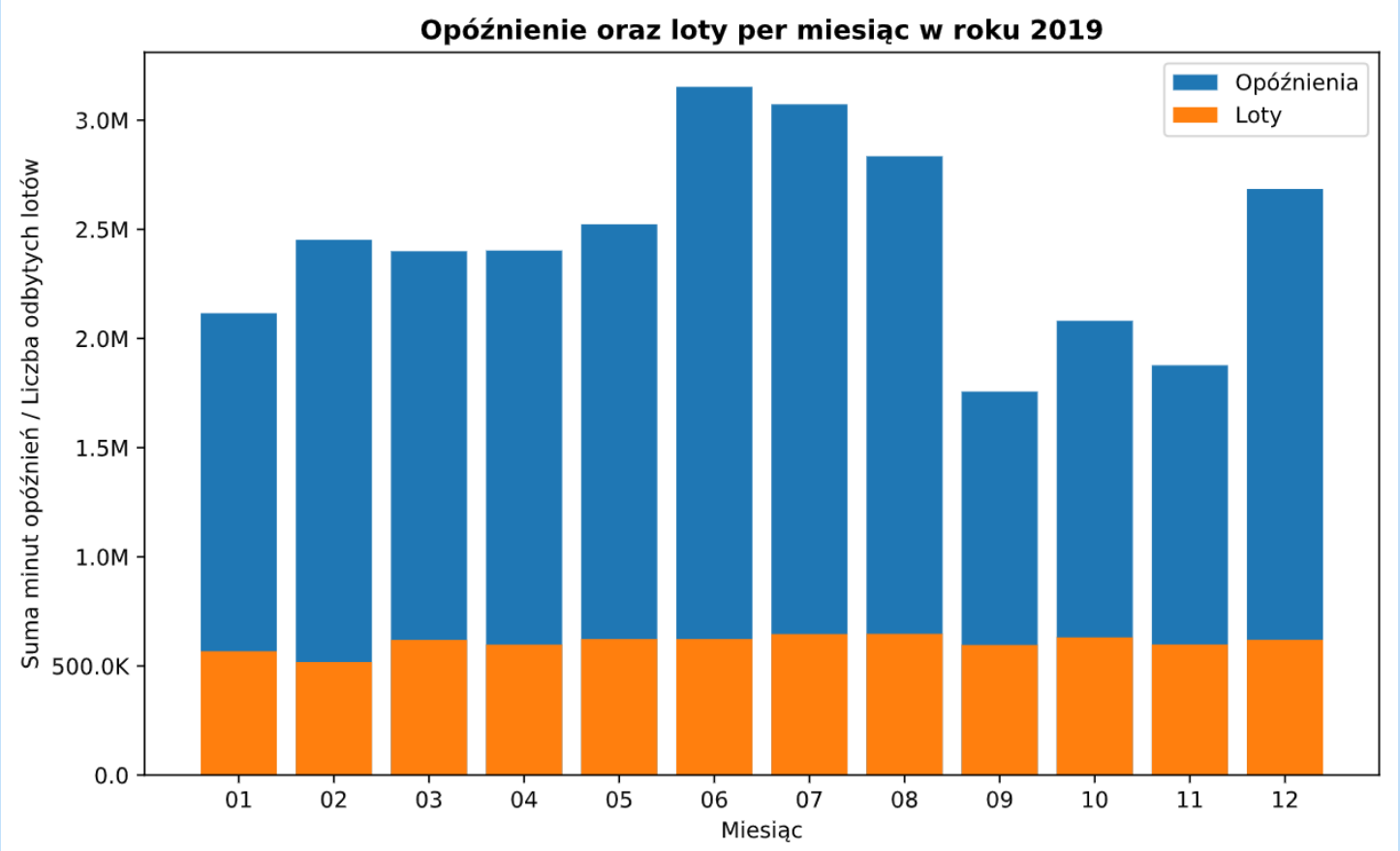

Opóźnienie per miesiąc w roku 2019
Suma minut opóźnień
3M
2M
2M
2M
1M
500K
0
01
02
03
04
05
06
07
08
09
10
11
12
Miesiąc


Opóźnienie każdego dnia w 2020
Suma minut opóźnień
140K
120K
100K
80K
60K
40K
20K
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
2020-11
Dzień

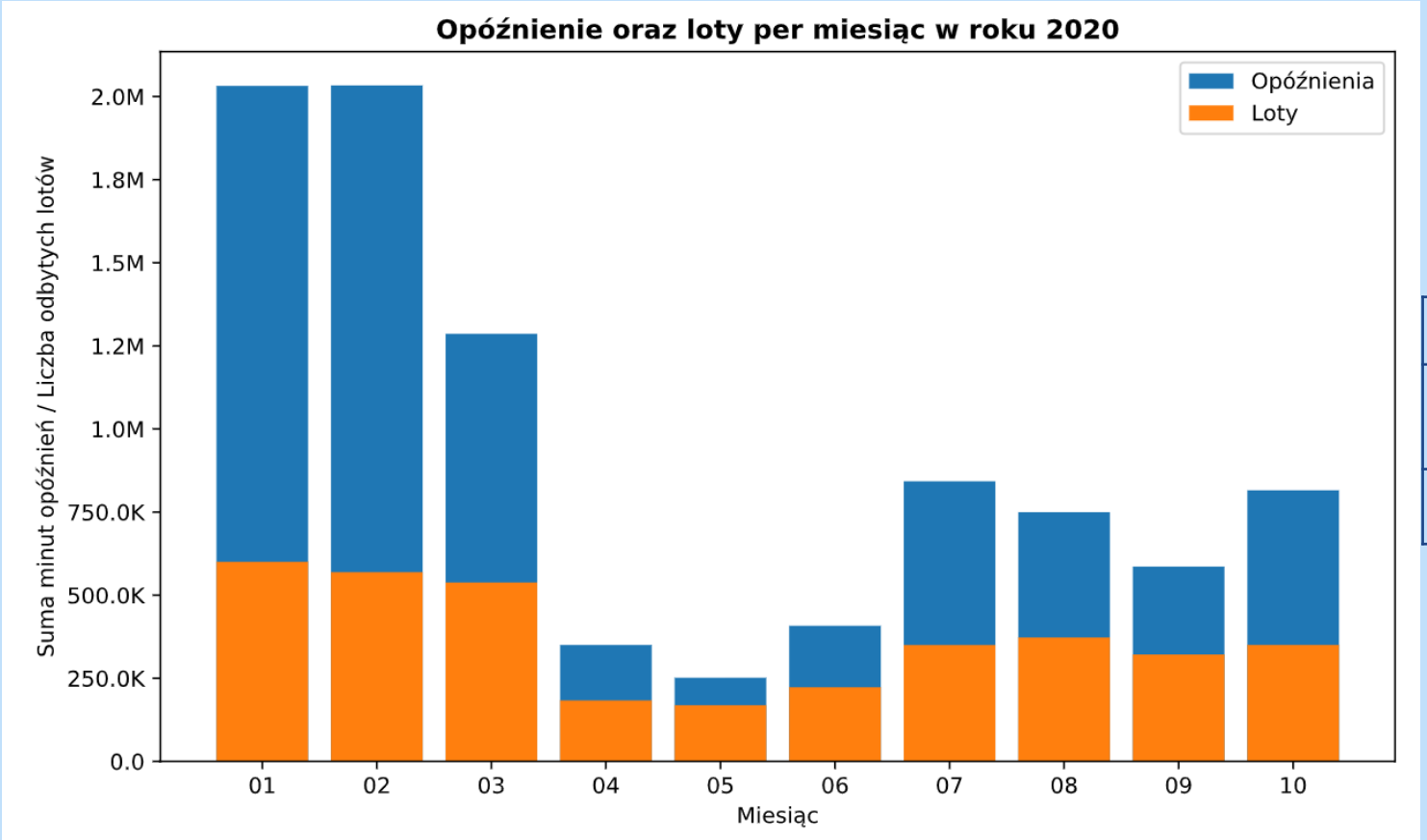

Opóźnienie per miesiąc w roku 2020
Suma minut opóźnień
2M
2M
2M
1M
1M
750K
500K
250K
0
01
02
03
04
05
06
07
08
09
10
Miesiąc


Opóźnienia danego miesiąca na przestrzeni lat 2009-2020
Suma minut opóźnień
25M
20M
15M
10M
5M
0
01
02
03
04
05
06
07
08
09
10
11
12
Miesiąc

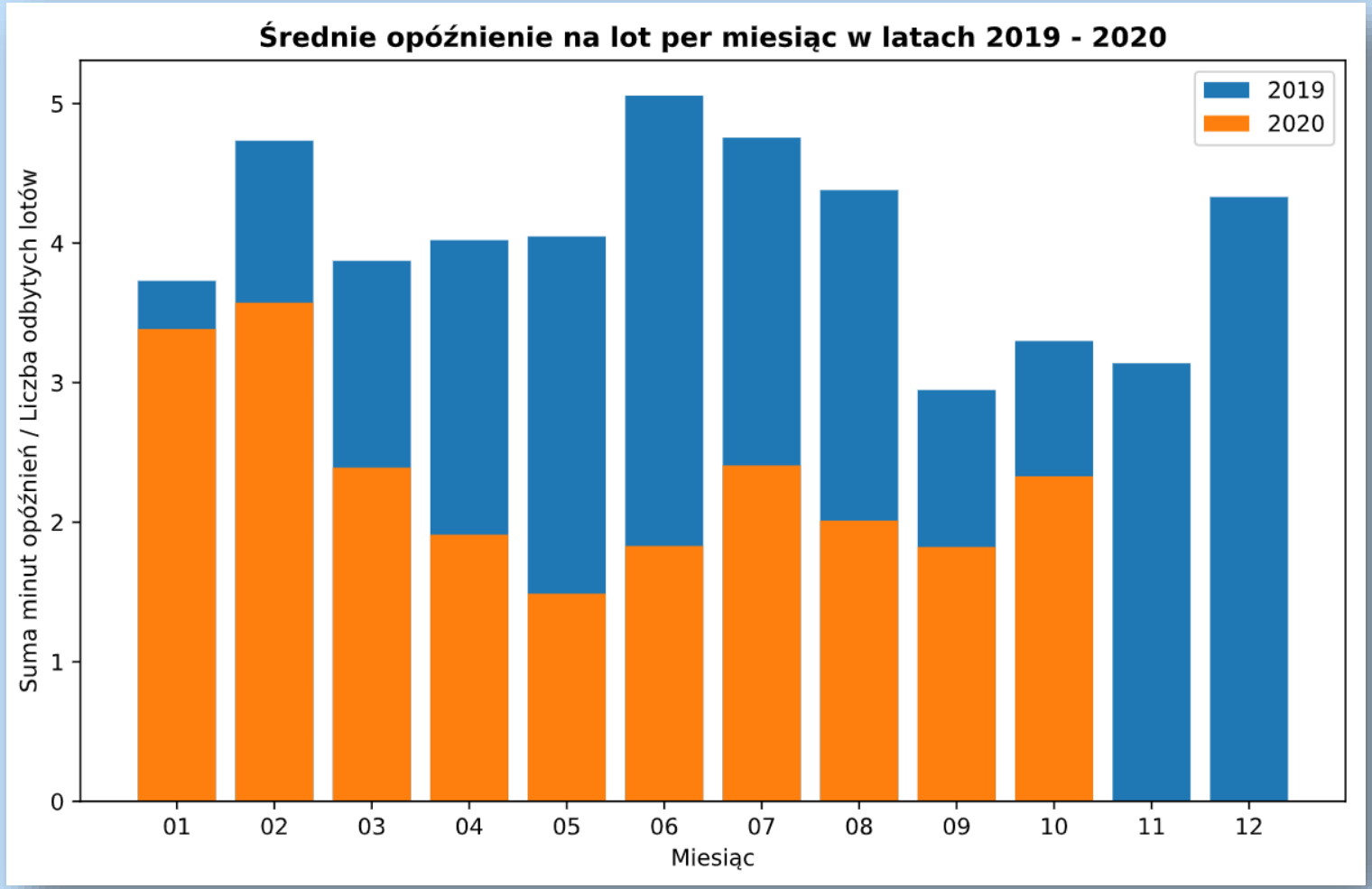

Opóźnienie per rok
30M
25M
20M
15M
10M
5M
0
Suma minut opóźnień
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Rok



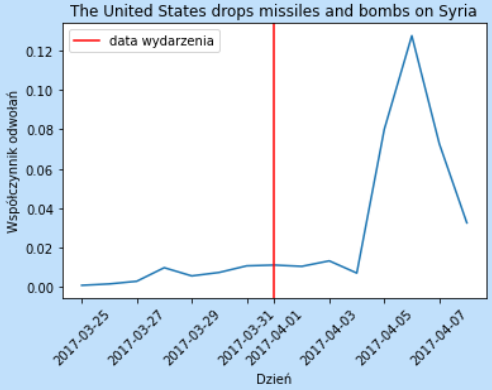
Pearson's r	**0.41**
Spearman's rho	**0.57**
Kendall's tau	**0.39**

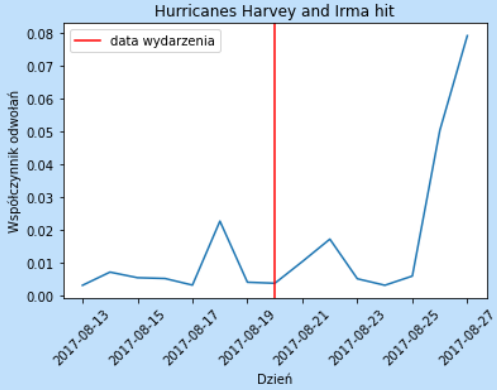


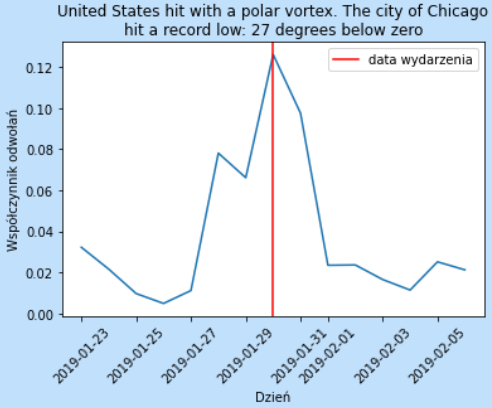
Pearson's r	**0.95**
Spearman's rho	**0.93**
Kendall's tau	**0.82**

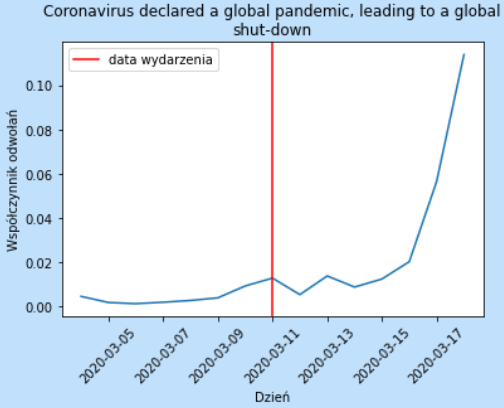

Średnie opóźnienie na lot per miesiąc w latach 2019 - 2020
2019
2020
Suma minut opóźnień / Liczba odbytych lotów
0
1
2
3
4
5
01
02
03
04
05
06
07
08
09
10
11
12
Miesiąc

Wydarzenia, które zachwiały stabilność branży

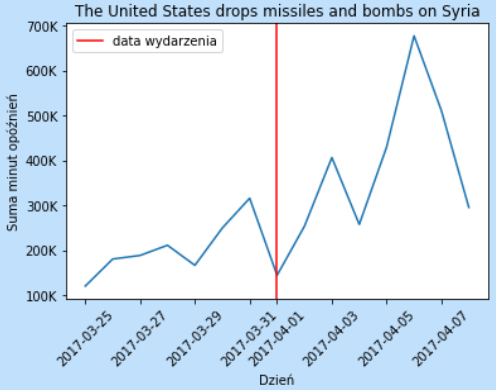


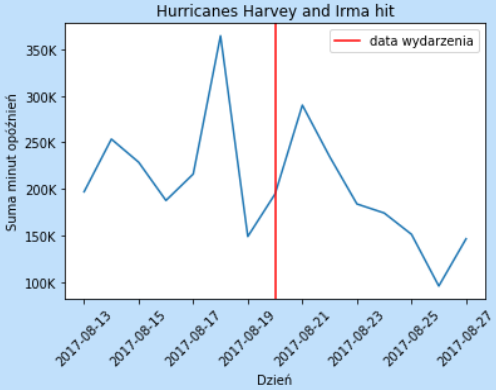


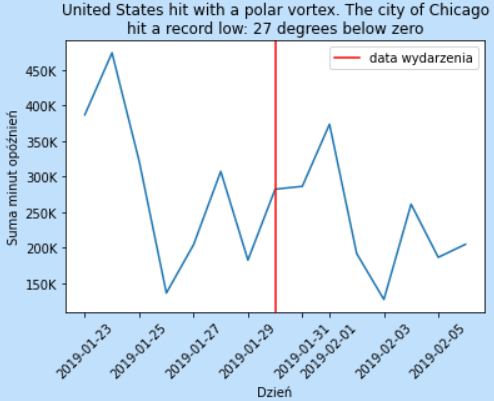


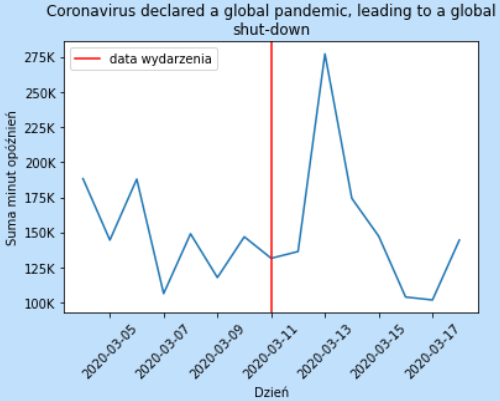


Wydarzenia, które zachwiały stabilność branży









BADANIA NAUKOWE

Klasteryzacja dla atrybutów dystansu i opóźnienia

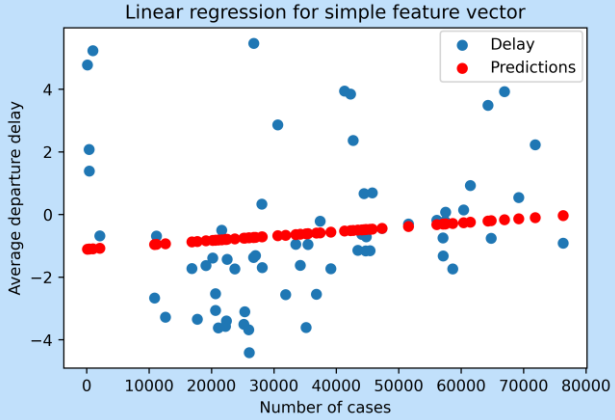
Współczynnik Silhouette dla kwadratu odległości euklidesowej

Liczba klastrów	KMeans	BisectingKMeans
5	0.724	0.599
2	0.823	0.823

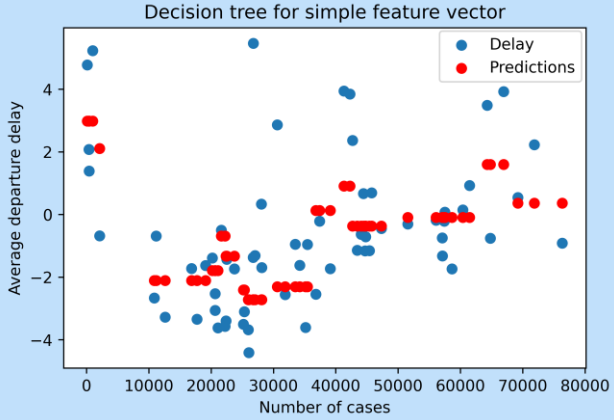
Macierz kowariancji w klastrach

Liczba klastrów	Kowariancja między atrybutami w klastrami
5	113.2, -117447.4, 824.9 -4553.4, -7809.7
2	439.8, 339.6

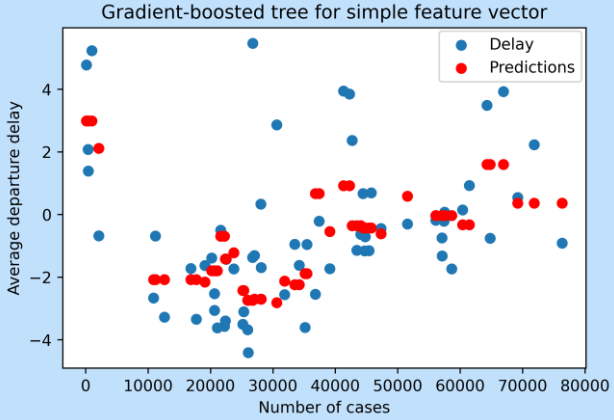
Regresja dla prostego wektora cech



RMSE = 2.31

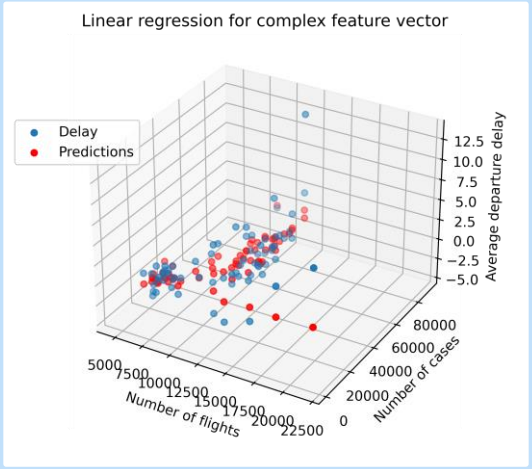


RMSE = 1.89

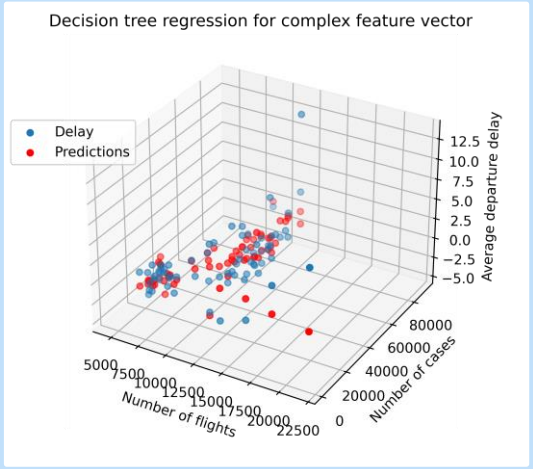


RMSE = 1.93

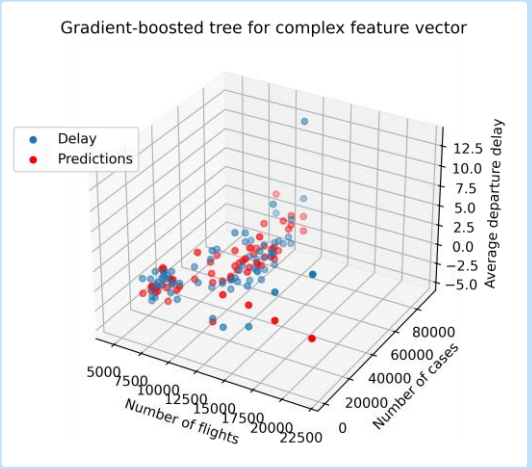
Regresja dla złożonego wektora cech



RMSE = 2.44



RMSE = 2.60



RMSE = 2.71

WNIOSKI

Wnioski biznesowe:

- Największe opóźnienia w okresach wakacyjnym i przedświątecznych
- Niewielki wpływ istotnych wydarzeń dekady przed 2020
- W latach przed 2020, liczba lotów nie wpływała na opóźnienie
- Najwięcej odwołanych lotów zarejestrowano na początku roku 2020 przez epidemię COVID
- Rozwój branży lotniczej postępujący od 2018 roku został gwałtownie zahamowany w 2020 roku i nie nastąpił powrót do poziomu sprzed 2018

Wnioski naukowe:

- Regresja metodą drzew decyzyjnych – obiecujący model predykcji opóźnienia na podstawie liczby zachorowań
- Wykorzystanie złożonego wektora cech wpłynęło negatywnie na wyniki regresji
- Przy podziale lotów na 2 klastry, współczynnik silhouette miał największą wartość, natomiast najmniejszą dla 5 klastrów miał dla algorytmu BisectingKMeans

DZIĘKUJEMY ZA UWAGĘ

BIBLIOGRAFIA

- Zbiory danych:

https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2018.csv
https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
https://covidtracking.com/
https://www.kaggle.com/gsnehaa21/federal-holidays-usa-19662020
https://en.wikipedia.org/wiki/Timeline_of_United_States_history_(1990%E2%80%932009)
https://en.wikipedia.org/wiki/Timeline_of_United_States_history_(2010%E2%80%93present)

- Materiały do prezentacji:

https://slidesgo.com/theme/aviation-consulting

- Repozytorium z kodem

https://github.com/jurczewski/PiADZD