

SQL - projekt - prezentacja

2018-03-11

Kurs Junior Data Scientist Zaoczne 1 (JDSZ1)

SQLuci

Monika Serkowska, Magdalena Kortas, Wojciech Artichowicz

1. **Czym się zajmujemy?**
2. **Diagram bazy danych**
3. **Raporty informacyjne**
4. **Raporty analityczne**
5. **Predykcje**

1.

Czym się zajmujemy jako grupa?

Czym się zajmujemy jako grupa?

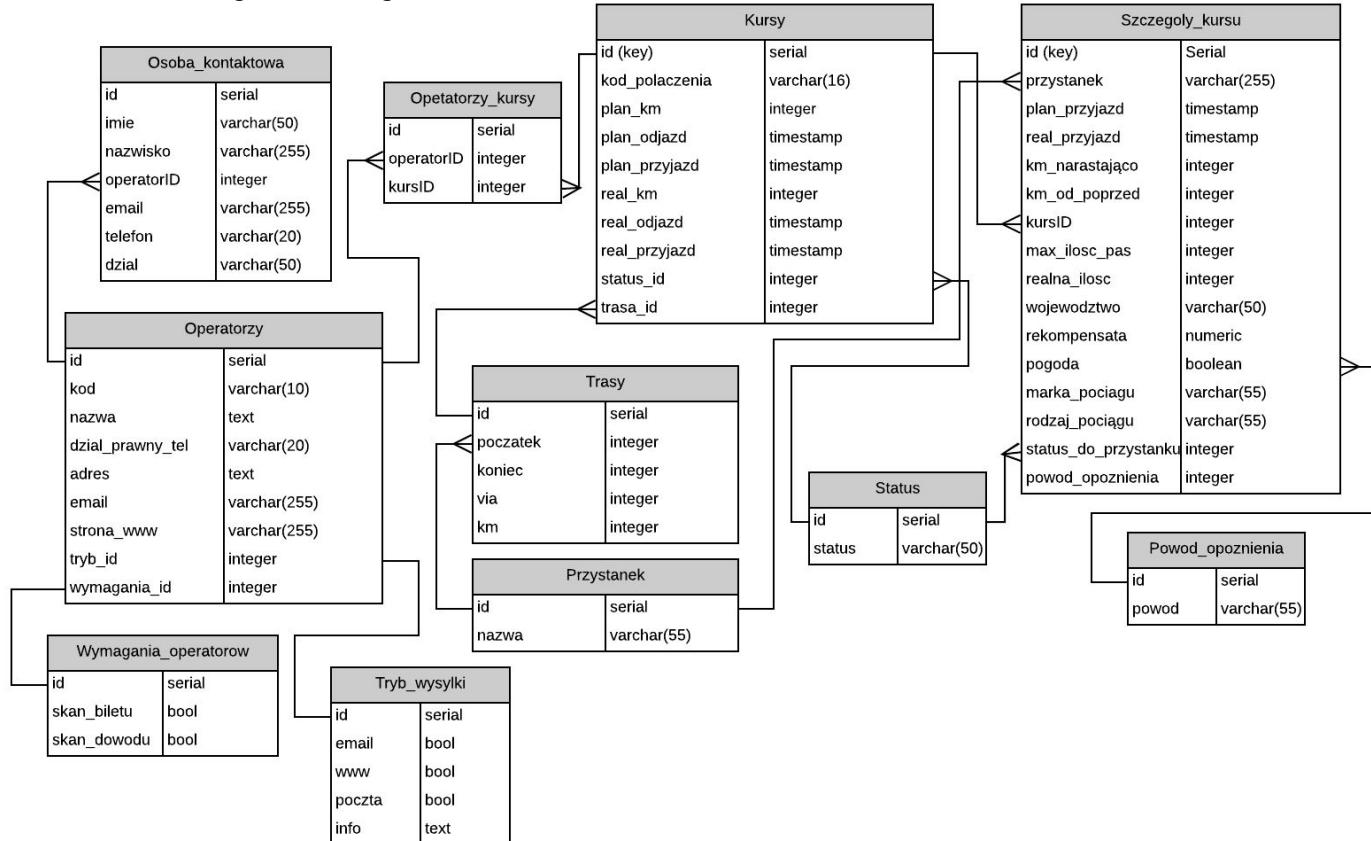
Analizujemy dane związane z **potencjałem tras, regionów, operatorów** czyli czynników wpływających na **opóźnienie** w celu **optymalizacji** pracy firmy i odnalezienia rozwiązań wpływających **pozytywnie na zysk**.

Stworzyliśmy 3 raporty informacyjne i 4 raporty analityczne.

Oprócz tego stworzyliśmy **predykcję** na marzec ilości zakłóconych tras, które podlegają odszkodowaniu.

2. Diagram bazy danych

Diagram bazy danych



3.

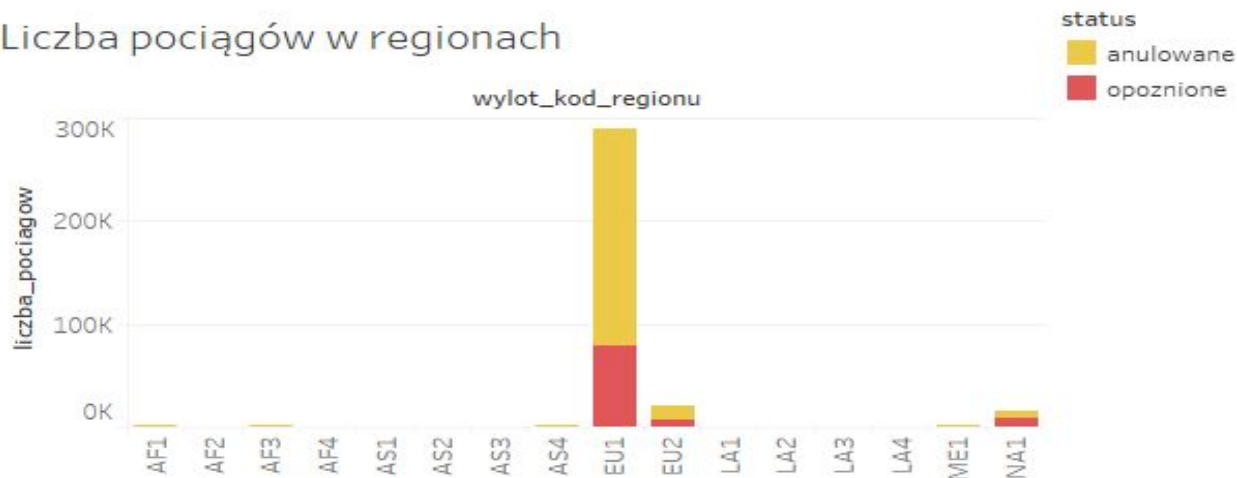
Raporty informacyjne

Potencjał regionów

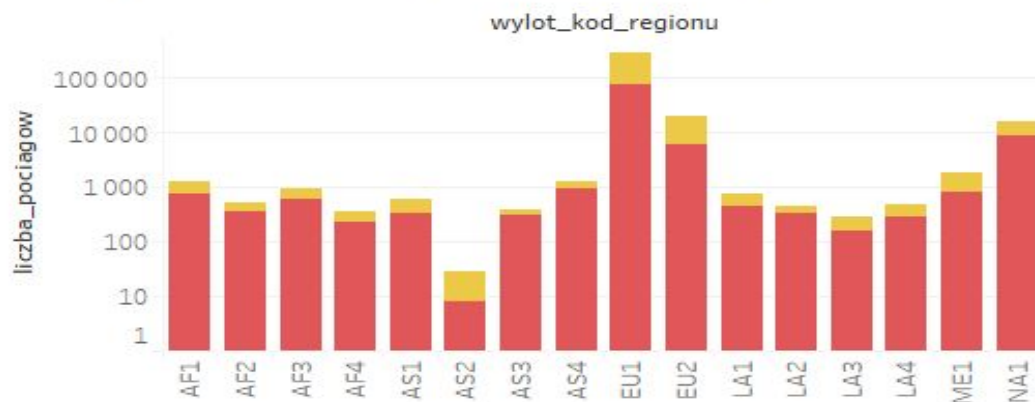
- **Cel:**
 - Który region przynosi/może przynieść największy zysk
- **Dla kogo:**
 - Informacje potrzebne do zaplanowania kampanii marketingowej
- **Impakt:**
 - klienci - większa dostępność informacji o możliwości uzyskania rekompensaty
 - zysk - ukierunkowanie kampanii na regiony o największym potencjale

Potencjał regionów

Liczba pociągów w regionach



Liczba pociągów w regionach - skala logarytmiczna



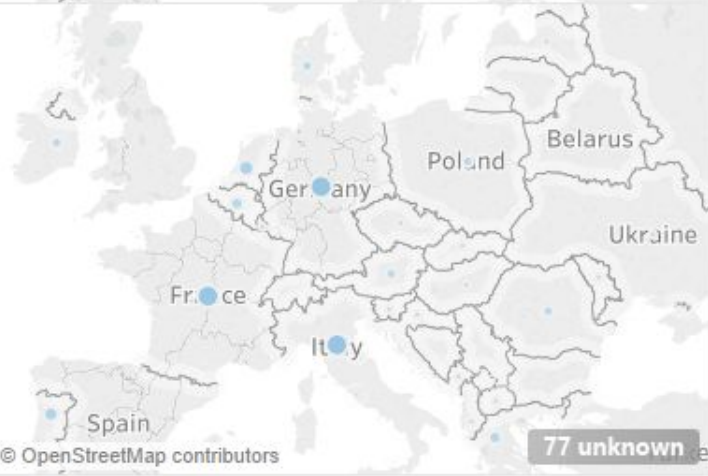
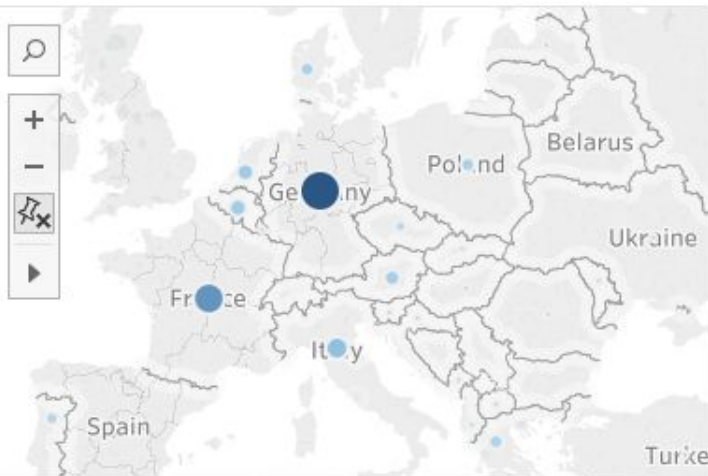
Potencjał regionów

Rekompensaty i liczba pociągów wg krajów

status

anulowane

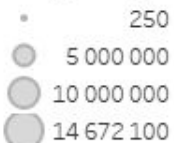
opoznione



© OpenStreetMap contributors

77 unknown

suma_rekompensat



liczba_pociagow



Potencjał operatorów

- **Cel:**
 - Który operator przynosi (może przynieść) największy zysk
- **Dla kogo:**
 - Informacje potrzebne do przygotowania kampanii marketingowej
- **Impakt:**
 - klienci - większa dostępność informacji o możliwości uzyskania rekompensaty
 - zysk - ukierunkowanie kampanii na pasażerów najsłabszych operatorów

Potencjał operatorów

Operator EiC

- 52,1 tys. tras odwołanych = kwota rekompensaty **15,41 mln**
- 22 tys. tras opóźnionych = **8,47 mln**
- Mamy najwięcej wniosków od pasażerów tego operatora

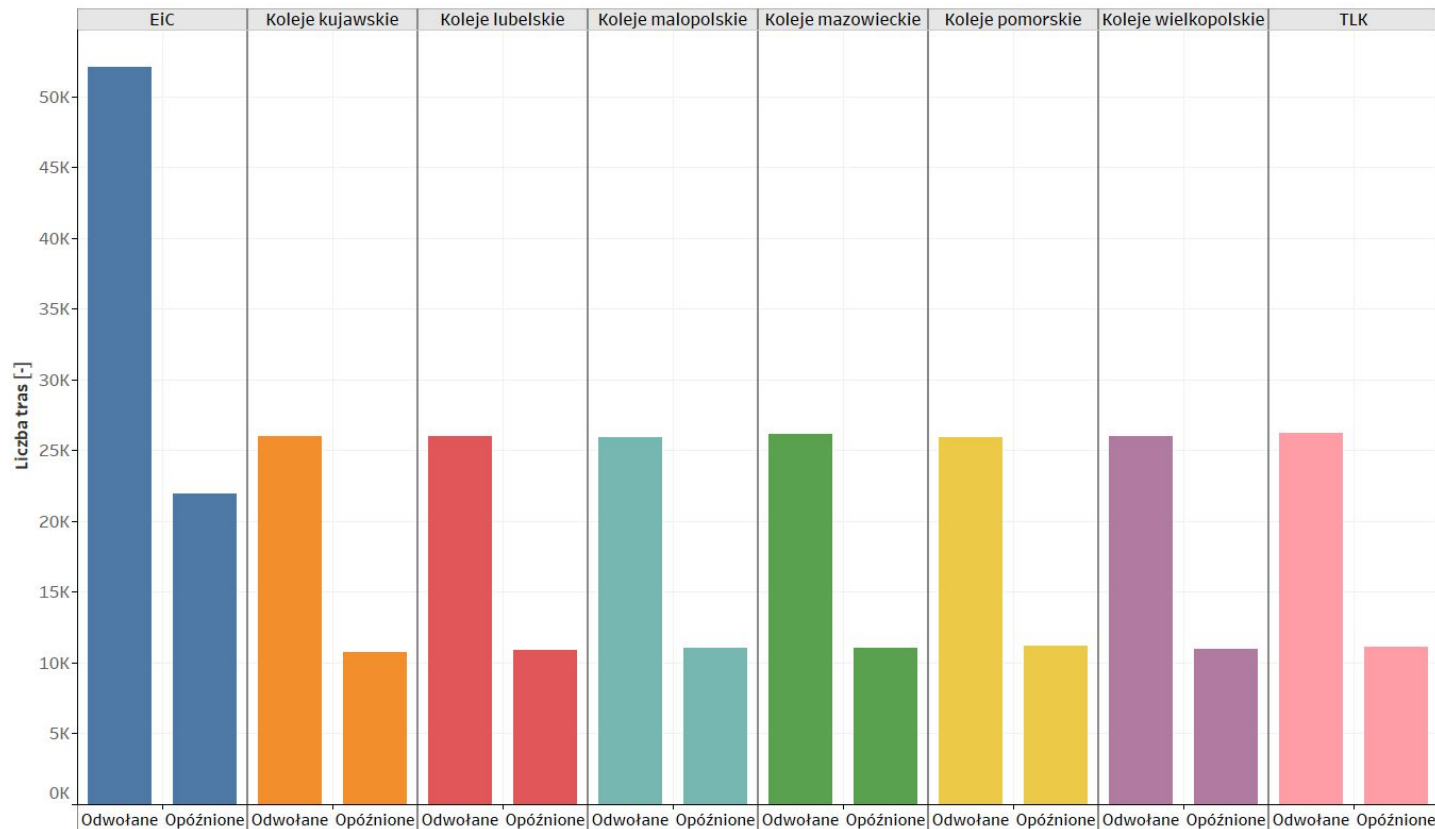
Pozostali operatorzy (średnia)

- 26 tys. tras odwołanych = kwota rekompensaty **7,7 mln**
- 11 tys. Tras opóźnionych = kwota rekompensaty **4,24 mln**

Średnia kwota rekompensaty za kurs opóźniony jest o 90 zł wyższa niż za kurs odwołany.

Potencjał operatorów

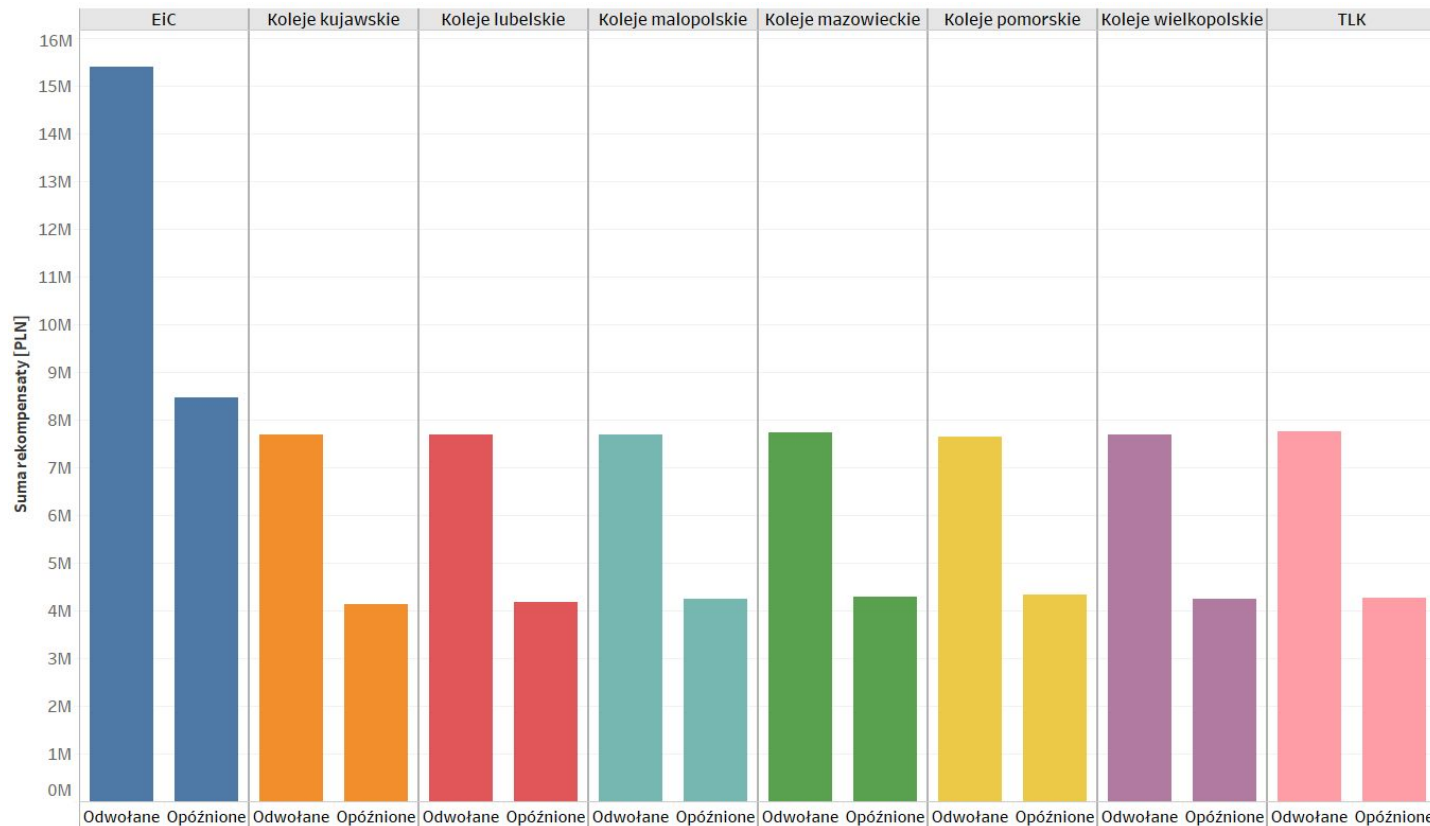
Liczba tras odwołanych i opóźnionych



(Wojtek)

Potencjał operatorów

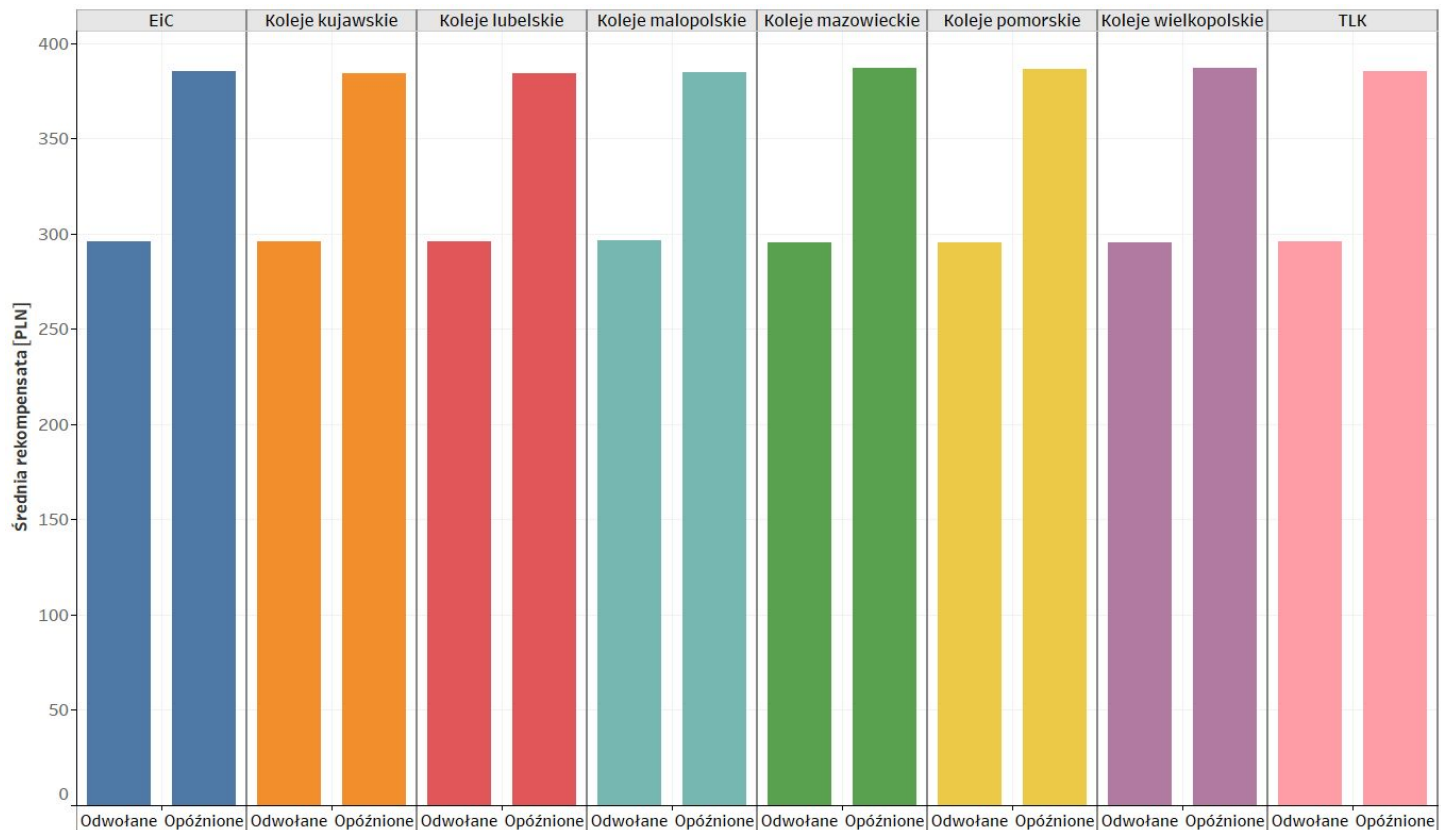
Całkowita kwota rekompensaty w badanym okresie



(Wojtek)

Potencjał operatorów

Średnia kwota rekompensaty



(Wojtek)

Potencjał operatorów

Sugestie

Dalsze analizy

- Badanie kosztu uzyskania rekompensaty
- Badanie czasu procesowania wniosku
- Badanie ilości odrzuconych / wygranych wniosków

Statystyki operatorów

- **Cel:**
 - Analiza operatorów (% **zakłóconych**, **ilość tras**, **pasażerów**) w celu wyznaczenia operatorów o największym potencjale jeśli chodzi o wielkość rekompensat
- **Dla kogo:**
 - Dla zarządu i działu marketingu, działu zajmującego się procesowaniem
- **Impakt:**
 - Pozytywny impakt na opinie klienta, ponieważ tej wiedzy jesteśmy w stanie łatwiej dotrzeć poprzez kampanię do poszkodowanych klientów danej firmy/operatora
 - Pozytywny impakt na czas procesowania, ponieważ można zrobić **ranking operatorów** i najpierw rozpatrywać wnioski, które statystycznie mają największą szansę na odszkodowanie
 - Pozytywny wpływ na zysk - **dotarcie do pasażerów** najczęściej zakłóconych tras

Statystyki operatorów

Statystyki Operatorów

operator	całkowita_rek..	ilosc_pas	ilosc_tras	procent_anulo..	procent_na_cz..	procent_opozn..	suma_pasaz_u..	procent_upra..
EIC	4 212 922 150	11 303 427	74 892	69,75%	0,00%	28,24%	11 304 301	100%
KKU	2 090 432 400	5 617 081	37 186	70,05%	0,00%	27,96%	5 617 081	100%
KLU	2 092 963 300	5 620 660	37 283	69,85%	0,00%	28,06%	5 620 784	100%
KMA	2 118 107 400	5 671 918	37 621	69,67%	0,00%	28,27%	5 672 188	100%
KML	2 115 559 450	5 668 108	37 386	69,50%	0,00%	28,44%	5 668 188	100%
KPO	2 118 216 550	5 674 815	37 539	69,14%	0,00%	28,82%	5 674 863	100%
KWI	2 114 092 200	5 669 612	37 461	69,59%	0,00%	28,30%	5 669 734	100%
TLK	2 116 846 800	5 679 493	37 712	69,62%	0,00%	28,37%	5 679 493	100%

4.

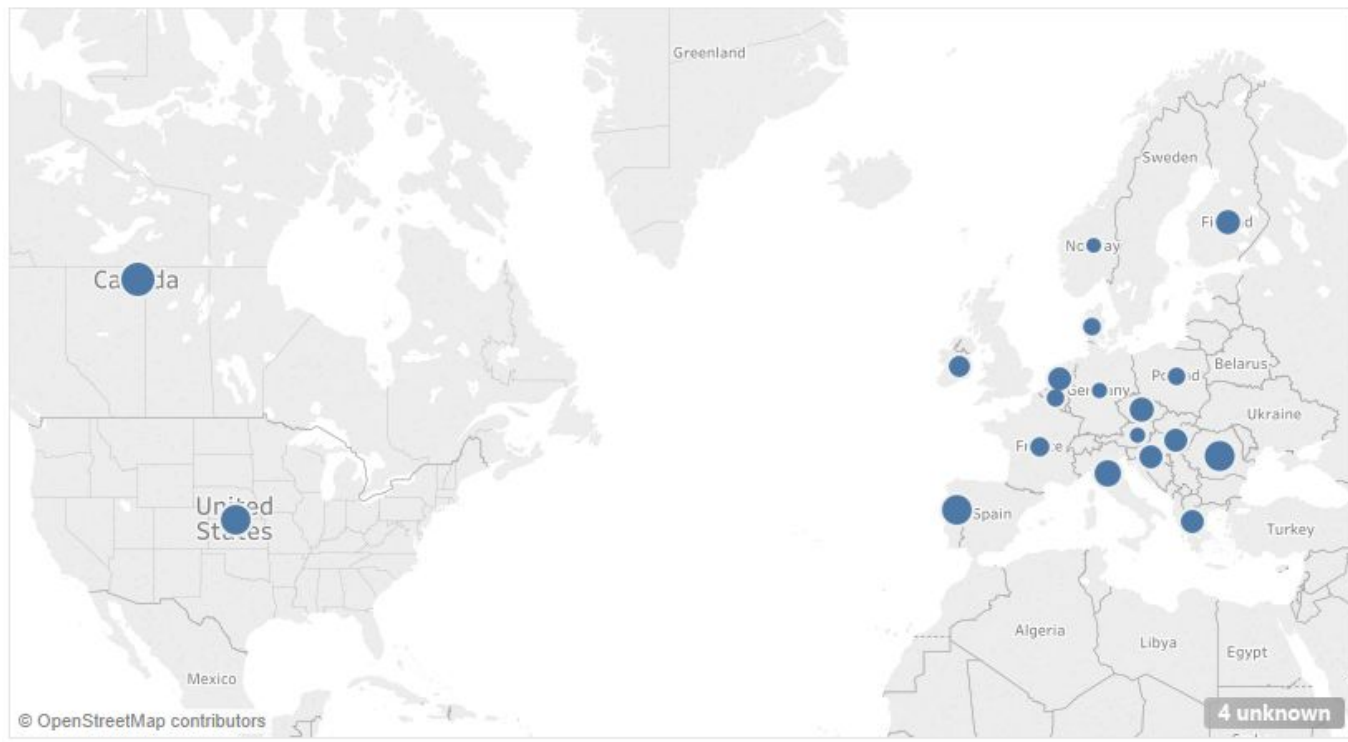
Raporty analityczne

Opóźnienia w regionach

- **Cel:**
 - Czy opóźnienia znacząco różnią się w regionach?
- **Dla kogo:**
 - Informacje potrzebne do przygotowania kampanii marketingowej
- **Impakt:**
 - na opinie klienta - pozytywny - większa dostępność informacji o możliwości uzyskania rekompensaty
 - na czas procesowania - pozytywny - wnioski z regionów o dużym potencjale mają pierwszeństwo w kolejce
 - na zysk - pozytywny wpływ - ukierunkowanie kampanii na kraje i lotniska o największym potencjale pozwoli lepiej wykorzystać środki na działania marketingowe

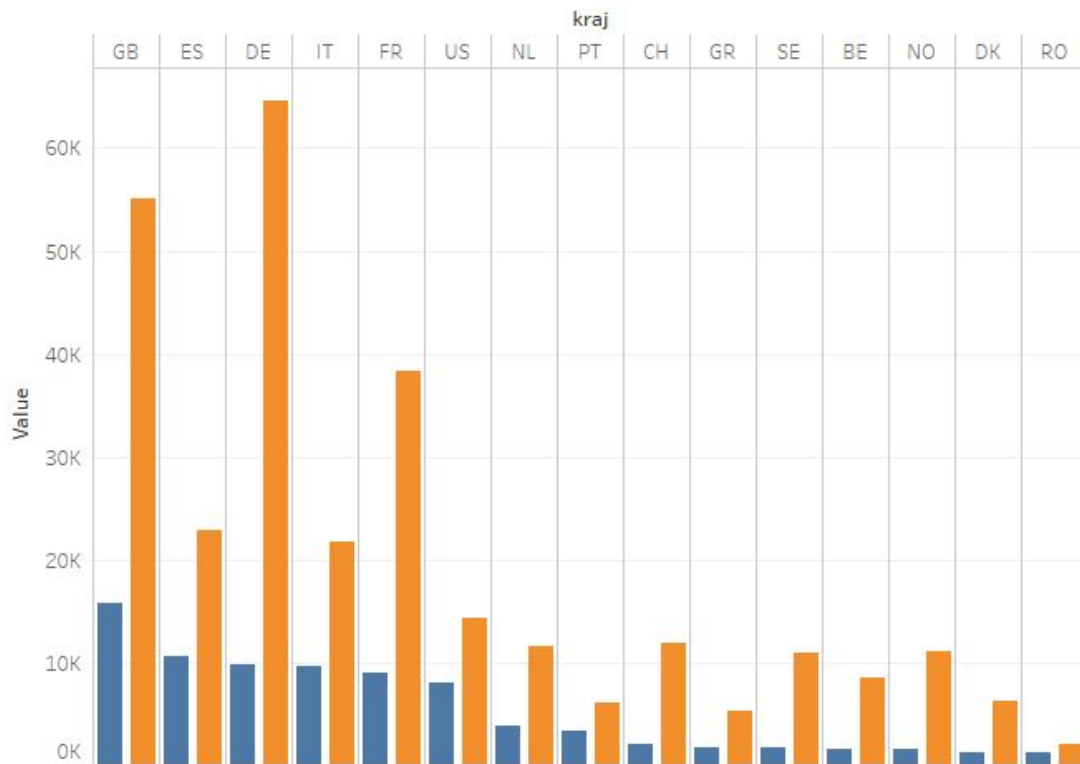
Opóźnienia w regionach

Procent opóźnionych lotów w krajach gdzie lotów było przynajmniej 1000



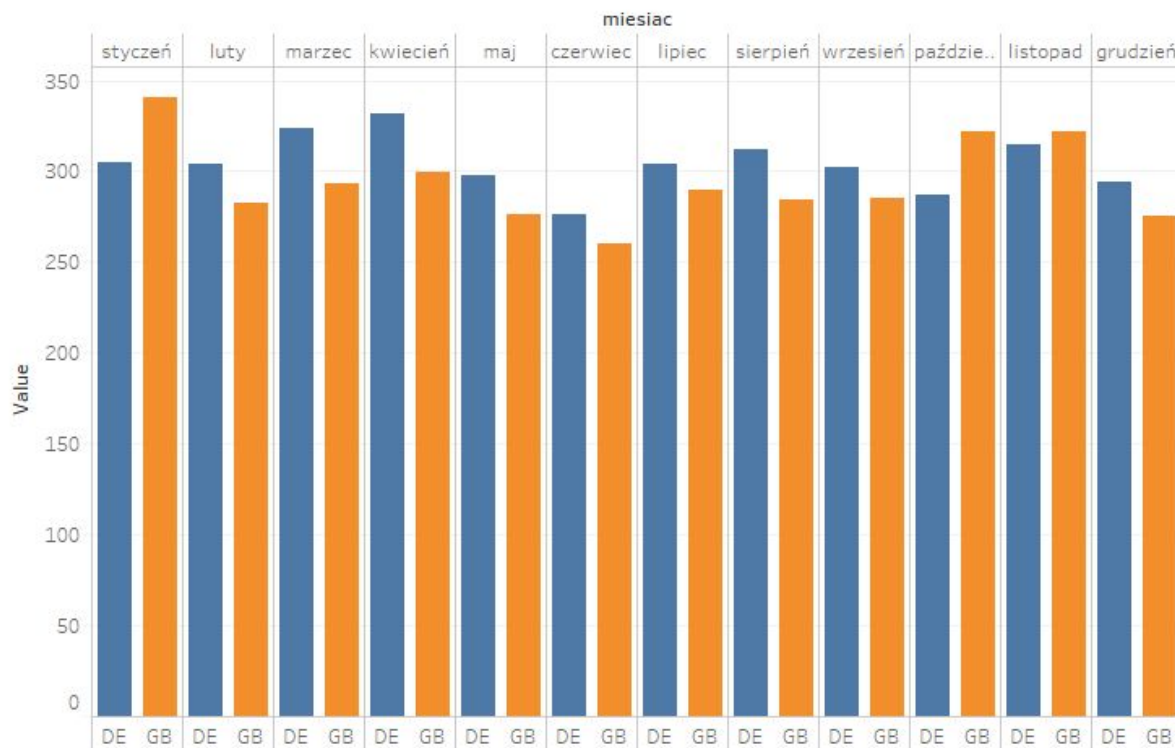
Opóźnienia w regionach

Liczba lotów ogółem i opóźnionych wg kraju



Opóźnienia w regionach

Średni czas opóźnienia



Najczęściej zakłócone trasy

- **Cel:**
 - Ranking najczęściej zakłóconych tras, które **podlegają odszkodowaniu** od linii lotniczej
- **Dla kogo**
 - Dla zarządu i działu marketingu, działu zajmującego się procesowaniem
- **Impakt**
 - Pozytywny impakt na opinie klienta, ponieważ dzięki znajomości najczęściej zakłóconych tras jesteśmy w stanie łatwiej dotrzeć poprzez kampanię do poszkodowanych osób
 - Pozytywny impakt na czas procesowania, ponieważ można zrobić **ranking połączeń** i najpierw rozpatrywać wnioski, które statystycznie mają największą szansę na odszkodowanie
 - Pozytywny wpływ na zysk - **dotarcie do pasażerów** najczęściej zakłóconych tras

Najczęściej zakłócone trasy

Top 10 Tras Opóźnionych 180 min

trasa	
BOS-FCO	99,69%
MIA-FCO	99,67%
JFK-FCO	97,02%
JFK-MXP	96,60%
ORD-LHR	55,93%
CDG-JFK	40,84%
JFK-LHR	39,75%
GLA-LHR	35,84%
LHR-JFK	35,30%
LHR-ORD	33,93%

Top 10 Tras anulowanych

trasa	
LUG-ZRH	100,00%
FRA-STR	100,00%
FRA-HAJ	99,69%
NUE-FRA	99,33%
FRA-NUE	99,22%
FRA-DUS	99,20%
DUS-FRA	98,86%
FRA-BRU	98,72%
MUC-DUS	98,68%
FRA-HAM	98,24%

Analiza wpływu wybranych zmiennych na opóźnienie

- **Cel:**
 - Uzyskanie ważnych zmiennych do modelu
 - Zmiana modeli kolejowania wniosków w oparciu o zmienne
- **Dla kogo**
 - Informacje potrzebne do przygotowania kampanii marketingowej
 - (Informacje potrzebne do usprawnienia procesu analizy wniosków)
- **Impakt**
 - Klienci - większa dostępność informacji w odpowiednich miejscach
 - Zysk - wyłowienie zmiennych wpływających na opóźnienie

Analiza wpływu wybranych zmiennych na opóźnienie

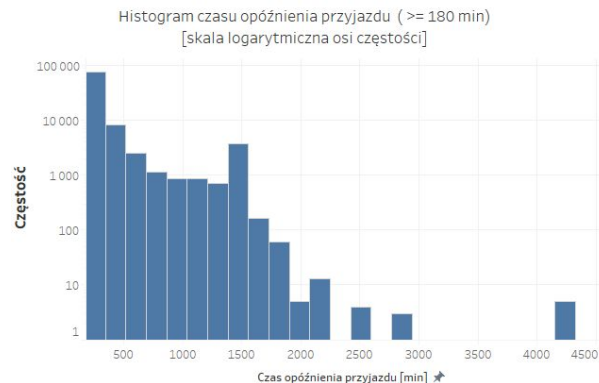
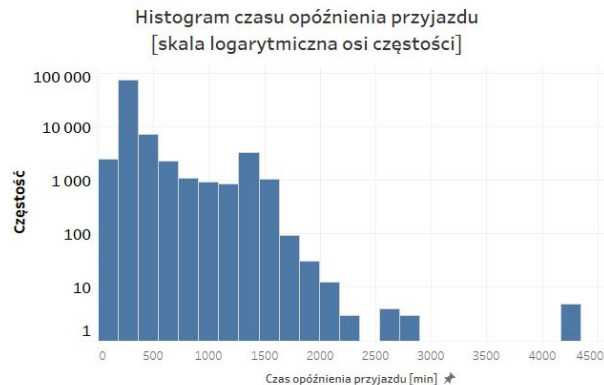
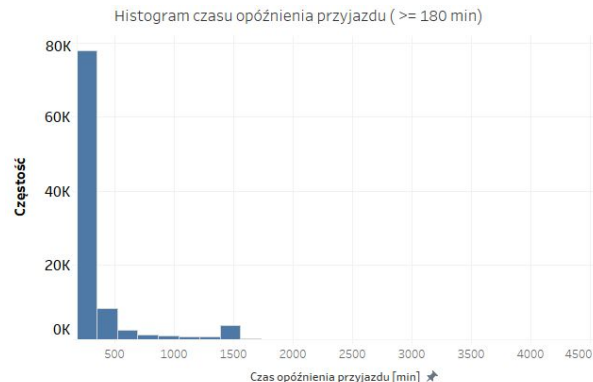
- Silny wpływ sprzętu na średni czas opóźnienia
- ~4% najbardziej opóźnionych wniosków realizowanych jest przez jeden typ sprzętu
- Niewystarczająca ilość danych do odpowiedzi na pytanie o sezonowość
- Wyraźnie inny przebieg w roku 2017
- Brak wpływu operatora na średni czas opóźnienia (zarówno względem sprzętu jak i czasu)

Podstawowe parametry

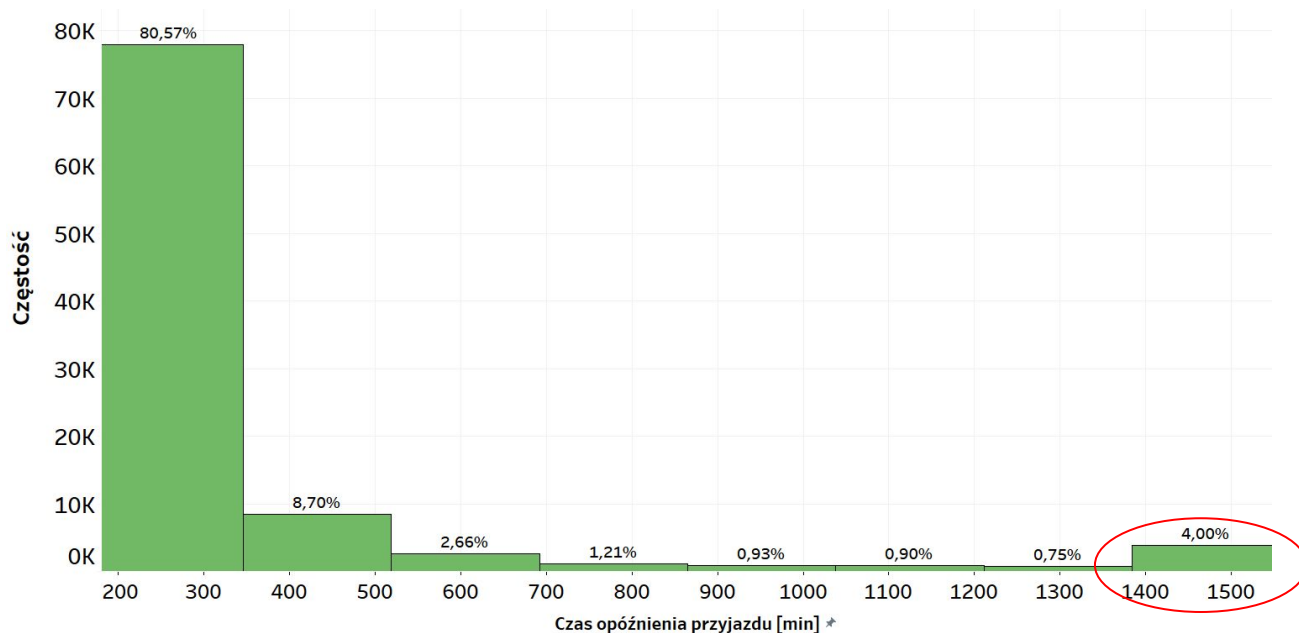
Nazwa	Czas opóźnienia [min]		
	Średnia	Mediana	Odchylenie Std.
EiC (EIC)	334	233	291
TLK (TLK)	335	234	289
Koleje pomorskie (KPO)	334	233	297
Koleje małopolskie (KML)	334	233	293
Koleje kujawskie (KKU)	335	232	293
Koleje lubelskie (KLU)	334	233	289
Koleje wielkopolskie (KWI)	334	234	293
Koleje mazowieckie (KMA)	335	233	293
Koleje śląskie (KSL)	null	null	null

Wynik testu
niezależności
operatora i
średniego czasu
opóźnienia **pv = 0,5**

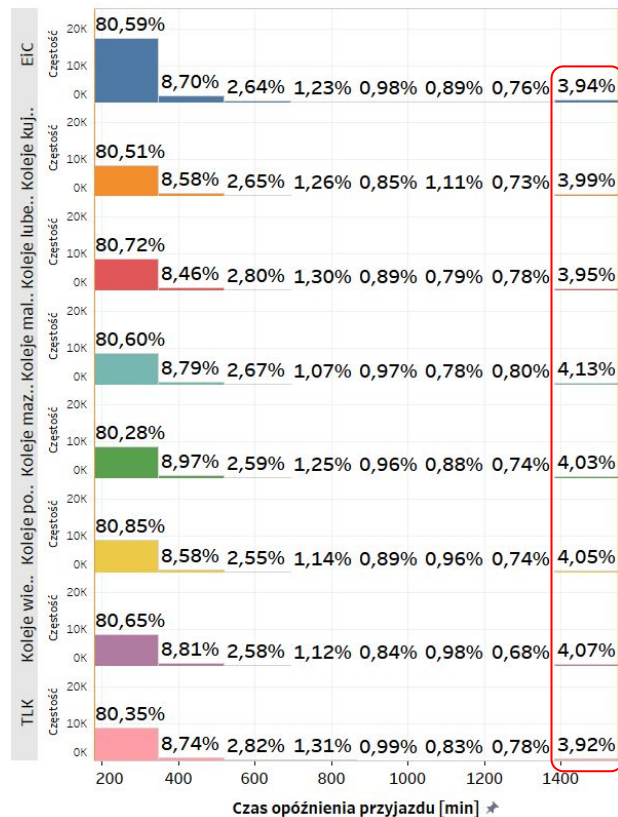
Analiza wpływu zmiennych na opóźnienie (histogram)



Analiza wpływu zmiennych na opóźnienie (histogram)



Analiza wpływu zmiennych na opóźnienie (histogram)

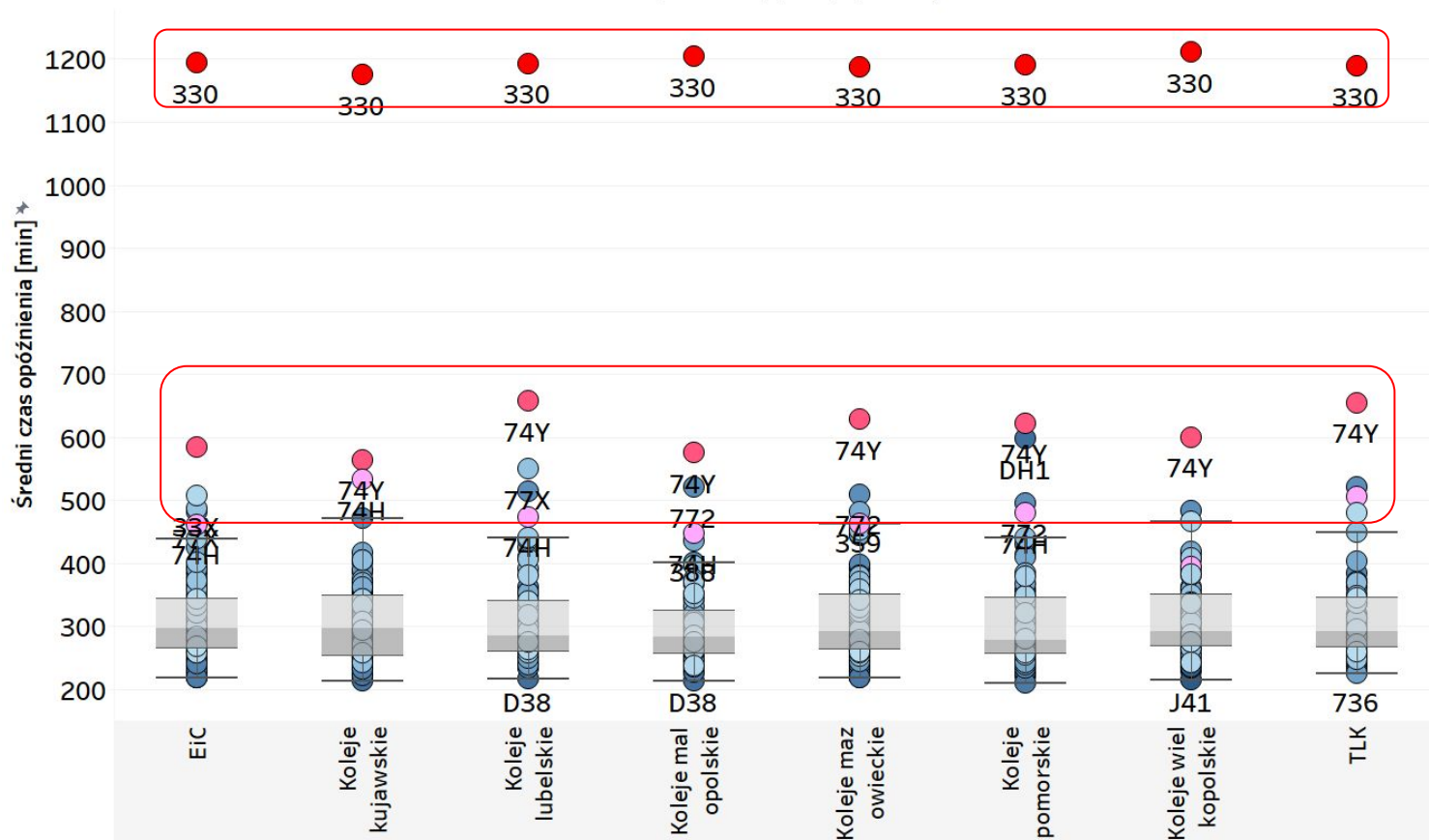


(Wojtek)

Wstępna analiza wpływu zmiennych na opóźnienie

Sprzęt / operator

Średni czas opóźnienia (sprzęt, operator)

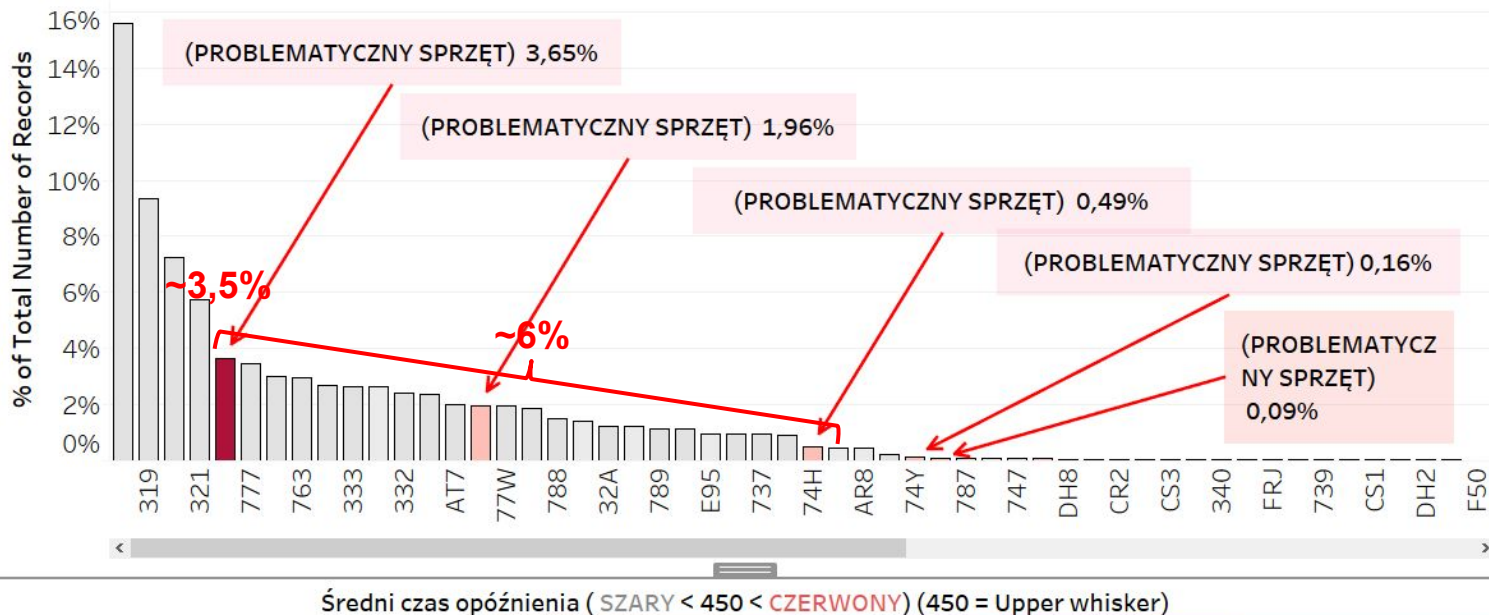


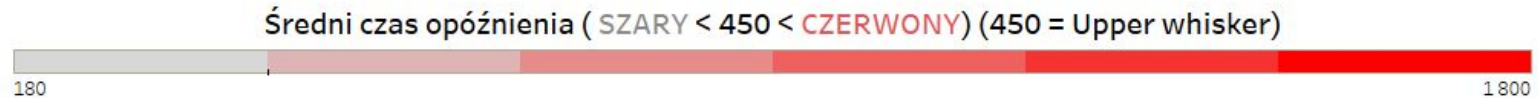
(Wojtek)

Wstępna analiza wpływu zmiennych na opóźnienie Sprzęt / operator

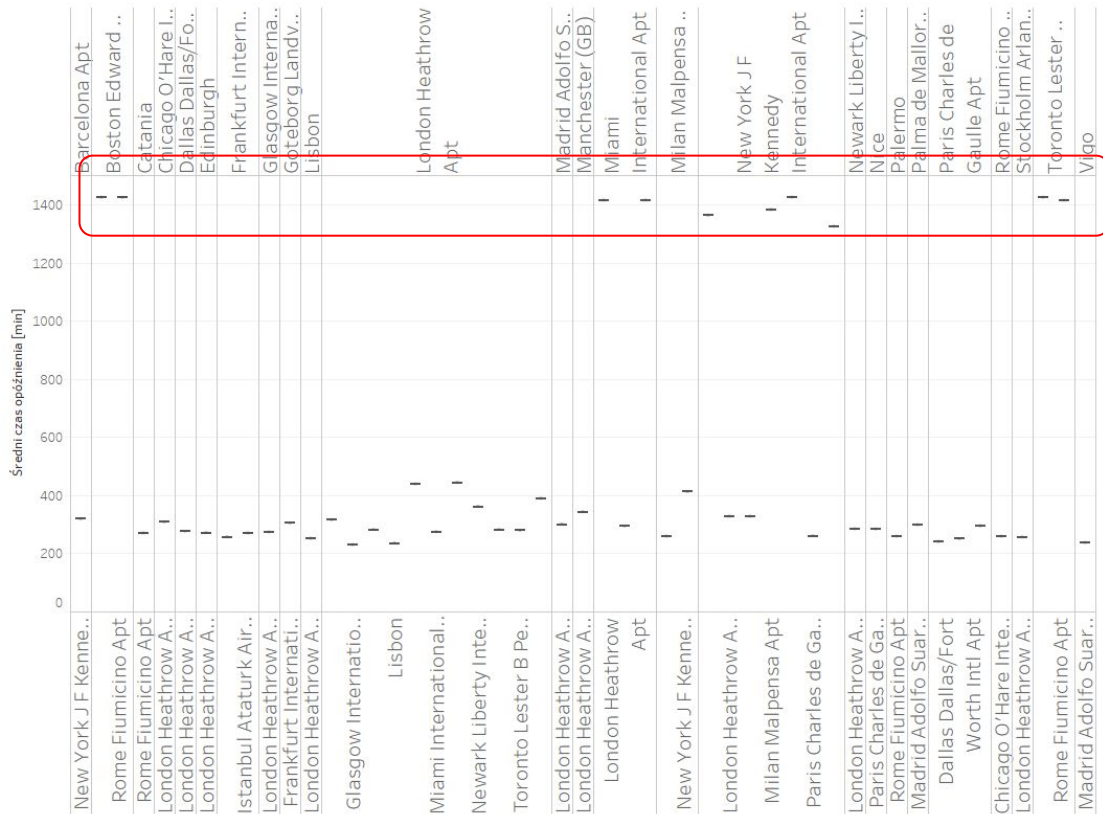
Wynik testu o
równości średniej w
grupie "330" i
reszcie sprzętu:
 6×10^{-1364}

Wynik testu o
równości średniej w
grupie sprzętu
problematicznego
bez "330" i reszcie
sprzętu:
 5.95×10^{-1376}

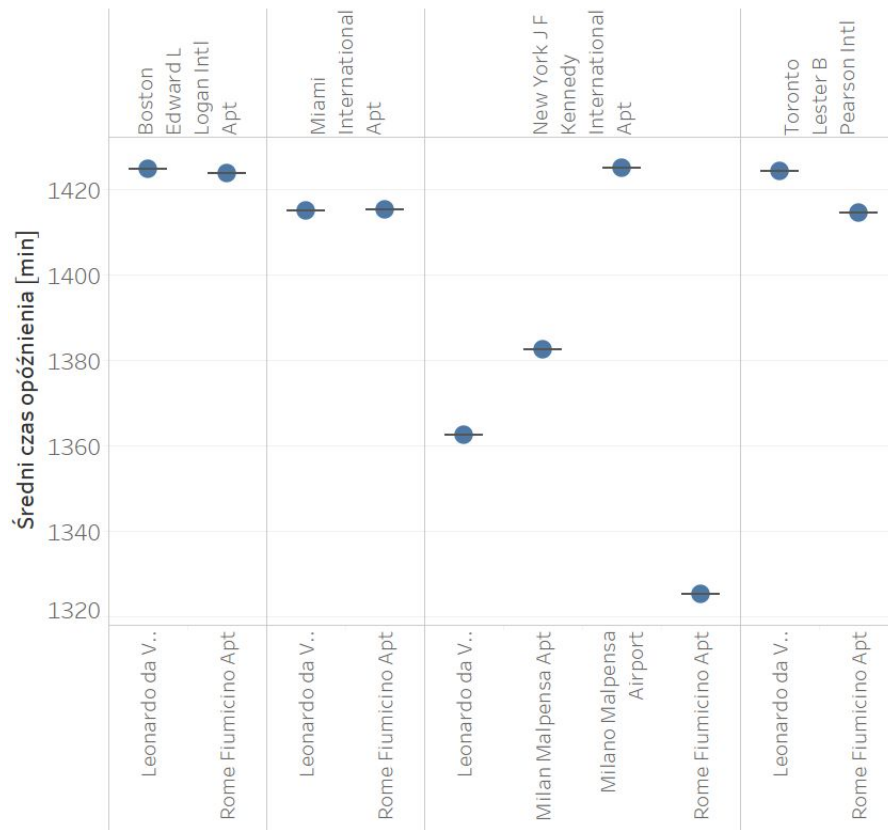




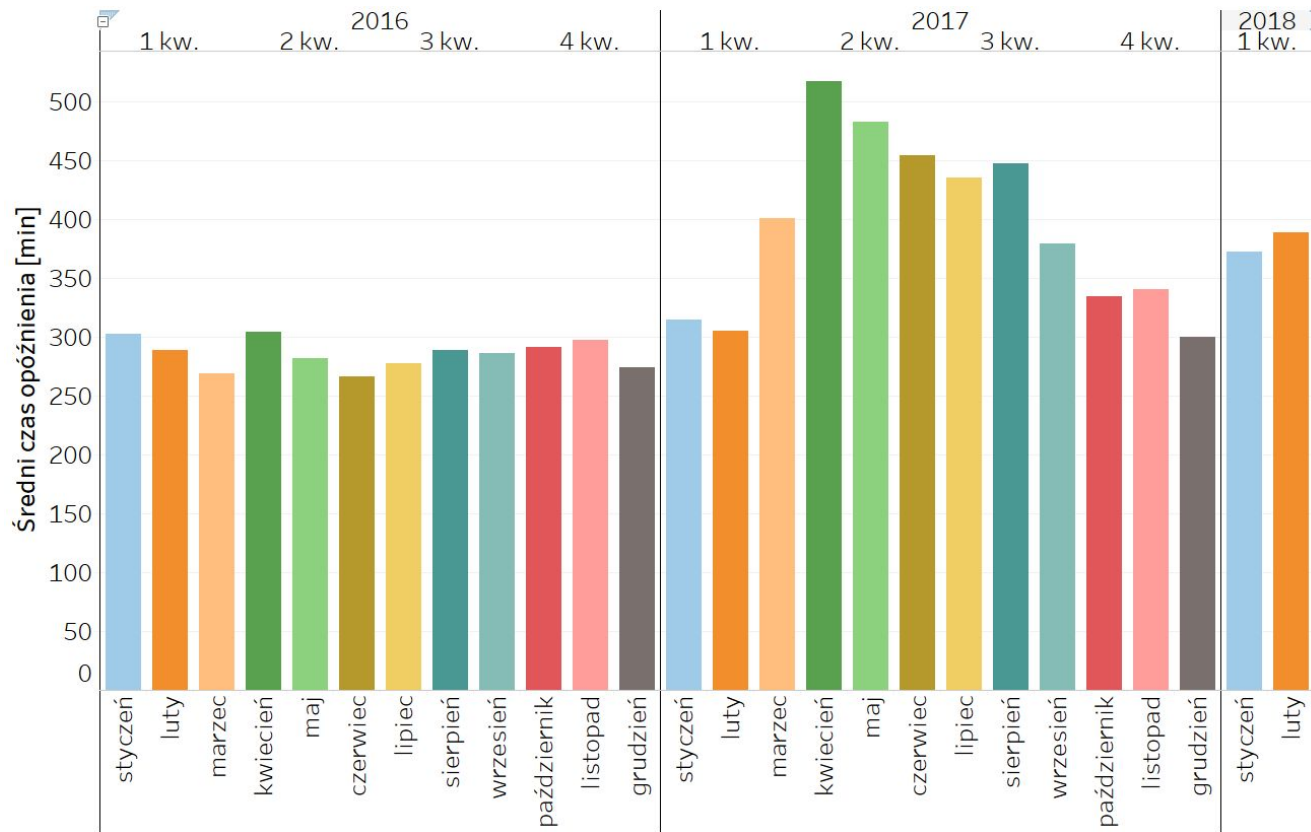
Średni czas opóźnienia wylot / przylot



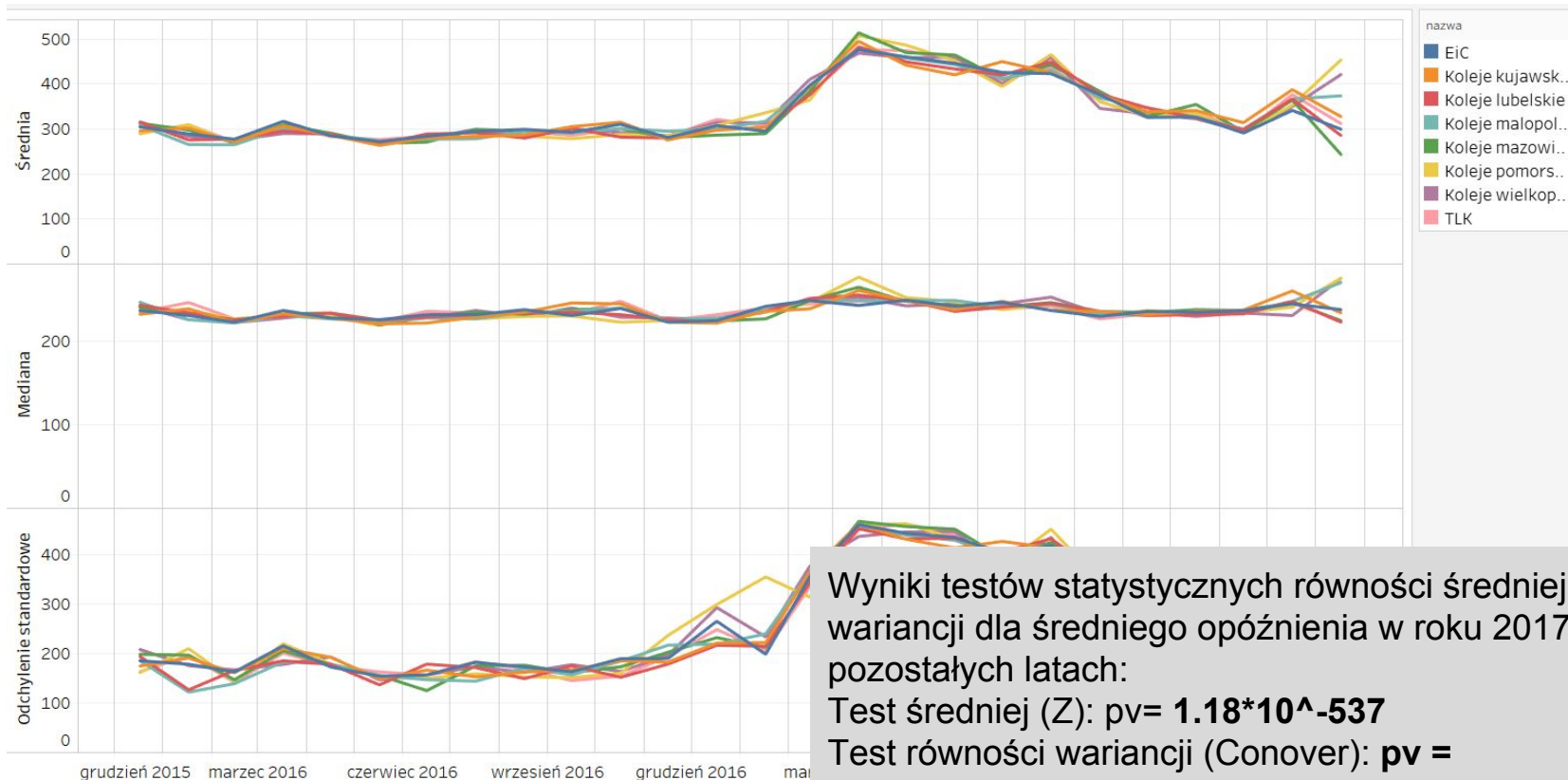
Średni czas opóźnienia wylot / przylot



Średni czas opóźnienia (miesiąc / kwartał)



Parametry opóźnienia (operator / rok)



Wyniki testów statystycznych równości średniej i wariancji dla średniego opóźnienia w roku 2017 i pozostałych latach:

Test średniej (Z): $p_v = 1.18 \cdot 10^{-537}$

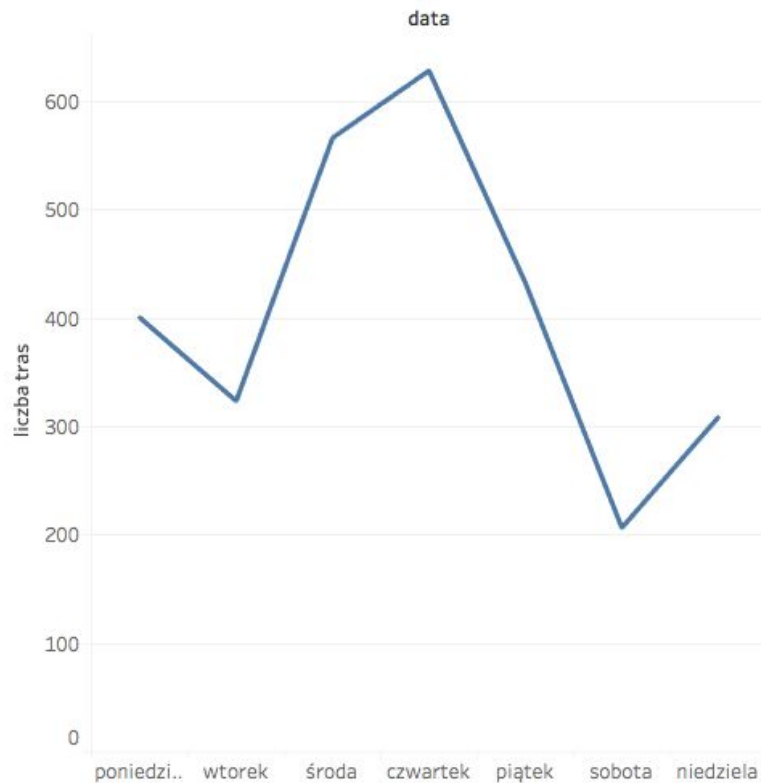
Test równości wariancji (Conover): $p_v = 2.83 \cdot 10^{-336}$

5. Predykcja

Predykcja

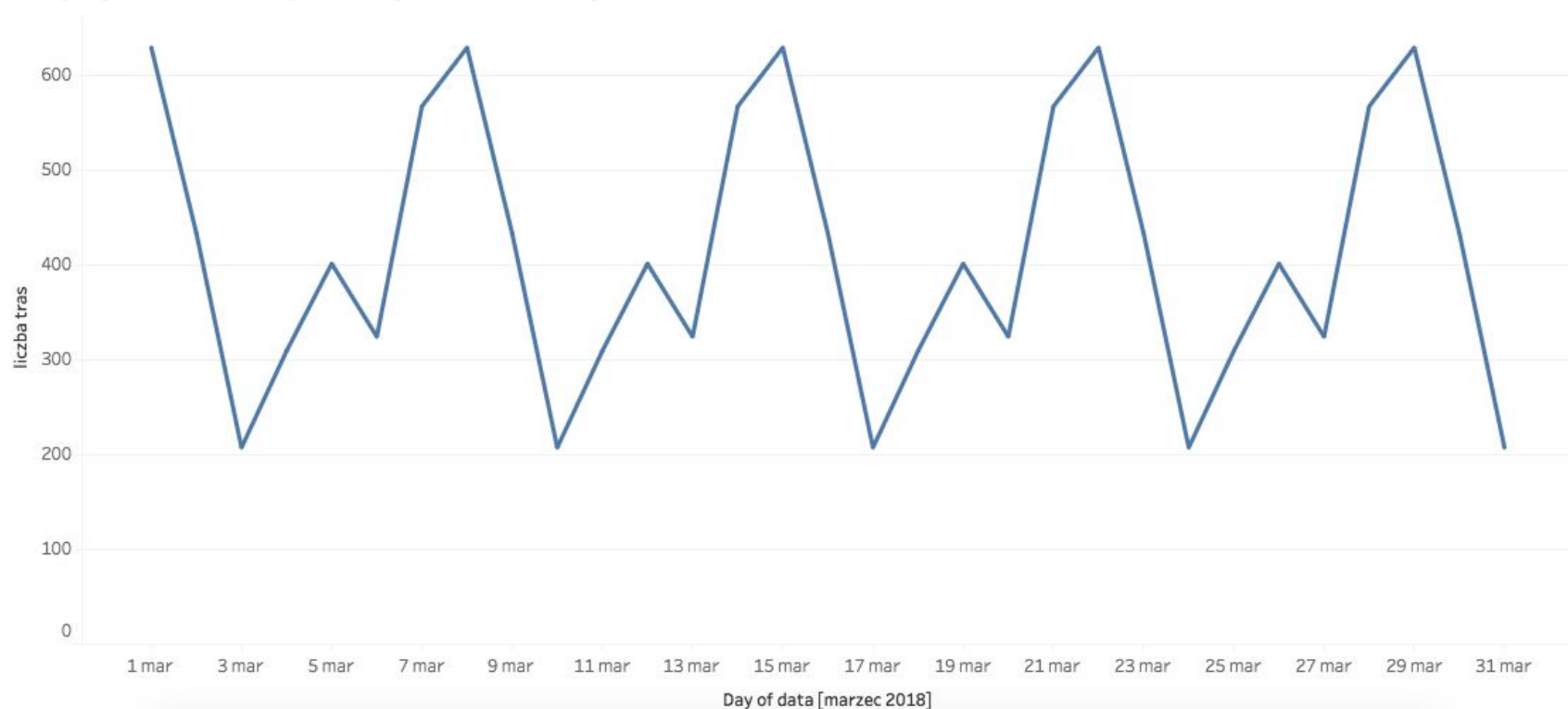
- **Cel:** Estymacja ilości tras zakłóconych na marzec, dla **przewidzenia ilości wniosków z szansą na pozytywne rozstrzygnięcie** a co za tym idzie, przychodu z prowizji oraz ilości pracy związanej z procesowaniem, by lepiej zaplanować czas pracy.
- **Dla kogo:** Dla zarządu, dla działu związanego z procesowaniem
- **Impakt:**
 - Pozytywny impakt na opinie klienta, ponieważ dzięki dobremu przewidzeniu ilości wniosków możemy **usprawnić proces analizy** - przyspieszyć go, więc klient szybciej otrzyma pieniądze
 - Pozytywny impakt na czas procesowania dzięki usprawnieniu procesu analizy
 - Brak wpływu na zysk

Predykcja tygodniowa



Predykcja na marzec

Predykcja ilości tras opóźnionych i anulowanych na marzec



Podsumowanie

- Promocja wśród klientów **linii EIC**, zwłaszcza na trasach najczęściej anulowanych w Niemczech, ale i opóźnionych (**USA-Włochy**).
- Najlepszy okres na kampanię na rynku **niemieckim** to **marzec, kwiecień i listopad**, kiedy występują najdłuższe opóźnienia.
- Stworzenie **systemu kolejkowego**, który by promował, jako pierwsze do analizy, wnioski, które mają największą **szansę na pozytywne rozpatrzenie**. Warto zwrócić uwagę na sprzęt (**wagony**), ponieważ mają one wpływ na opóźnienia. Pomoże to zaplanować i **przyśpieszyć** proces analizy, wiedząc też, że najwięcej wniosków wpływa w **czwartki** i pozytywnie wpłynie na **opinie klientów**.

Cel kampanii: dotarcie do nowych 1000 osób miesięcznie, z średnią prowizją o wysokości 100 euro = **1,2 miliona euro dodatkowego zysku rocznie**

Podsumowanie raportów - całkowity impact

Raport	Impact na opinie klienta	Impact na czas procesowania	Impact na przychód
Analiza operatorów, regionów i tras	pozytywny	pozytywny	1,2 miliona euro / rok
Wpływ wybranych zmiennych na opóźnienie	pozytywny	pozytywny	brak
Predykcja ilości wniosków na marzec	pozytywny	pozytywny	brak
Razem			1,2 miliona euro / rok

Podsumowanie - współpraca w grupie

- Stworzyliśmy wspólnie schemat bazy danych ERD, utworzyliśmy tabele. Po zasileniu danymi przeprowadziliśmy wspólnie eksplorację danych
- Podzieliliśmy po równo zadania dotyczące raportów informacyjnych, analitycznych i predykcji (każda osoba otrzymała 3 zadania)
- Podczas pracy nad zadaniami używaliśmy Jiry
- Komunikowaliśmy się przez Slack
- Mieliśmy jedno spotkanie robocze w ciągu tygodnia w O4, przedyskutowaliśmy wspólnie efekty pracy i pracowaliśmy nad impaktem każdego z raportów
- Wspólnie pracowaliśmy nad prezentacją



Linki

- **Jira:**
 - [Nasz board](#)
- **Github**
 - [Nasze repozytorium](#)
- **Tableau** - linki do naszych analiz:
 - [Analizy operatorów](#)
 - [Potencjał regionów](#)
 - [Liczby według krajów](#)
 - [Średni czas opóźnienia](#)
 - [Wpływ zmiennych](#)



Dziękujemy

Pytania?
Slack / email