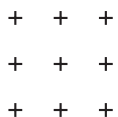




Wrocław — 30.05.2018

Analiza aktywności wypożyczalni samochodów elektrycznych Vozilla



DATA SCIENCE W PRAKTYCE



Informację o planowanym uruchomieniu we Wrocławiu wypożyczalni samochodów elektrycznych przyjąłem z wielką radością - w końcu każdy, kto czytał relacje z testowania limuzyn Tesli, marzył o własnej jeździe próbnej. Nie przeszkadzało mi, że pod marką Vozilla dostępne będą Nissany Leaf, kompaktowy elektryk to nadal elektryk. Z wypożyczalni skorzystałem w pierwszych godzinach jej funkcjonowania. Aplikacja mobilna pobrała opłatę za kurs, a ja - zachwycony nowym doświadczeniem - zadumałem się: ile jeszcze osób skorzysta dziś z tego pojazdu?



”

Zapraszam do lektury raportu zawierającego analizę aktywności wypożyczalni Vozilla. Jest on w całości oparty o dane, które systematycznie i przez dłuższy czas pobieraliśmy z API serwisu webowego.

Tomasz Zieliński
Team Leader — PGS Software S.A.



Gromadzenie i czyszczenie danych.

Ten rozdział opisuje techniczne aspekty obróbki danych. Jeśli Cię to nie interesuje, przeskocz do następnego rozdziału.

Witryna i aplikacja mobilna Vozilli prezentują na mapie miasta lokalizację aut dostępnych do wypożyczenia, wraz z ich szacowanym zasięgiem i procentowym stanem naładowania akumulatora. Dane te są odświeżane na bieżąco - wystarczy jedno proste żądanie HTTP, aby uzyskać plik danych w formacie JSON z zestawem informacji o całej flocie.

Rejestracja takich danych jest prosta. W systemach uniksowych (np. Linux) mamy do dyspozycji polecenie cron, które pozwala uruchamiać wybrany program w zadanych odstępach czasu. W opisywanym przypadku programem tym jest curl, łączący się ze wskazanym adresem www i zapisujący pobraną treść do pliku. Zleciłem pobieranie co minutę danych o wszystkich samochodach Vozilli, począwszy od pierwszego dnia komercyjnego działania usługi. Sprawdziłem, że cały mechanizm działa, pozostało uzbroić się w cierpliwość i czekać.

Sześć miesięcy później miałem do dyspozycji

ponad 252 000 plików zajmujących na dysku przeszło 13 gigabajtów

Zgodnie z oczekiwaniami dane nie były kompletne - czasem strajkowały serwery Vozilli, innym razem problemy z komunikacją zaliczała moja maszyna - w sumie jednak braki stanowiły mniej niż pół procenta całego zbioru. Innymi słowy kilkanaście brakujących godzin na pół roku gromadzenia danych.

Połowę braków stanowiły pojedyncze utracone pliki, np. trzy losowe minuty spośród 1440 składających się na kompletny zapis dobowy. Małe luki zostały

„zaspachlowane” danymi z sąsiednich pozycji. Pojedyncza brakująca minuta była uzupełniana danymi z minuty poprzedzającej, dwie brakujące minuty były uzupełniane danymi z minut otaczających lukę. Jedyne ryzyko przy takim podejściu to sklejenie dwóch wypożyczeń w jedno, bo w około 10% przypadków auta są rezerwowane lub wypożyczane w czasie krótszym niż trzy minuty. Pozostałe luki mają zazwyczaj od kilku do kilkunastu minut. **Analiza obejmuje dane ze 177 dni i nocy regularnego zapisu stanu floty.** Niestety, wskutek błędu ludzkiego utracony został cały zapis z 31 grudnia 2017, niemniej łączna liczba brakujących danych nie przekracza 1% ogółu.

Rzut oka na dane.

Każdy przetwarzany plik zawiera następujący zestaw informacji o samochodach Vozilli:

```
▼ 0:
  discriminator: "vehicle"
  platesNumber: "DW9R181"
  sideNumber: "181"
  color: "Biały"
  type: "CAR"
  ▼ picture:
    id: "72cfa8d5-14f9-48d7-9417-e72b780a8ae4"
    name: "72cfa8d5-14f9-48d7-9417-e72b780a8ae4"
    extension: null
  rangeKm: 158
  batteryLevelPct: 85
  reservationEnd: null
  status: "AVAILABLE"
  locationDescription: null
  id: "e009f35e-a5d7-488c-bcc2-6ec51534c3d3"
  name: "NISSAN Leaf"
  description: null
  ▼ location:
    latitude: 51.127391
    longitude: 16.934675
  metadata: null
```

Niektóre z pól były zawsze puste (np. metadata albo description), inne były nadmiarowe (np. jednakowe trójki „id-platesNumber-sideNumber”), jeszcze inne homogeniczne (wszystkie auta miały typ „vehicle”).

Do dalszej obróbki pobierzemy więc następujące zmienne:

- numer boczny
- rodzaj pojazdu (osobowy/van)
- szacowany zasięg w kilometrach
- naładowanie baterii w procentach
- limit czasu trwania rezerwacji
- bieżący status
- długość geograficzna
- szerokość geograficzna
- data i godzina zapisu informacji

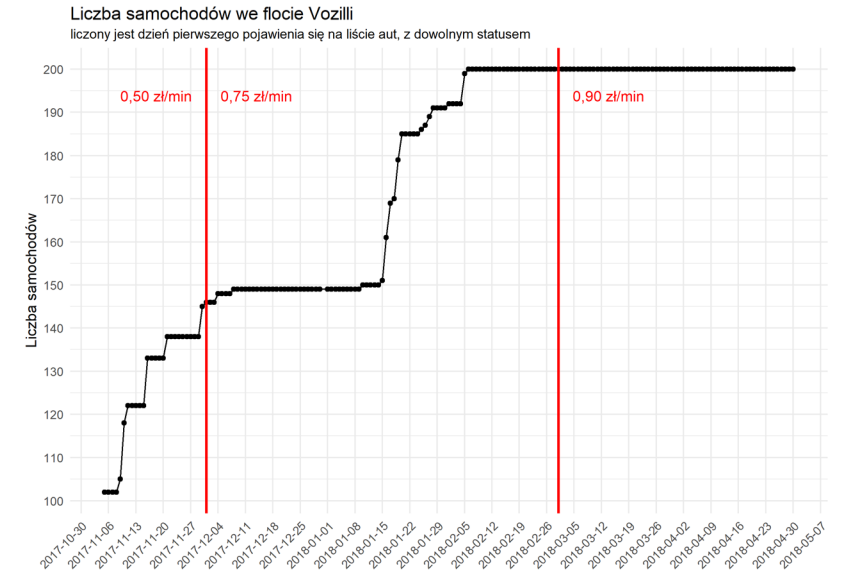
Wszystkie analizy i wykresy zostały sporządzone w języku R. Początkowa ramka danych składała się wyłącznie z wymienionych dziewięciu zmiennych (kolumn), które wystąpiły w 30.510.555 obserwacjach (wierszach).

Jesteś zainteresowany surowymi danymi do samodzielnej analizy? Namiary na githuba znajdziesz na końcu tekstu.

Analiza danych.

Wszystkie wykresy zostały wygenerowane przy użyciu języka R i pakietu ggplot2. Link do skryptu generującego komplet ilustracji znajdziesz w dodatku nr 3.

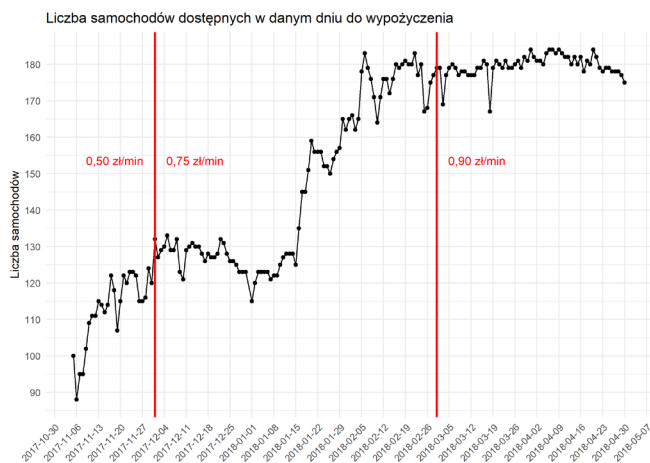
Pierwsze pytanie: Ile Nissanów dostępnych jest w sieci Vozilla?



”

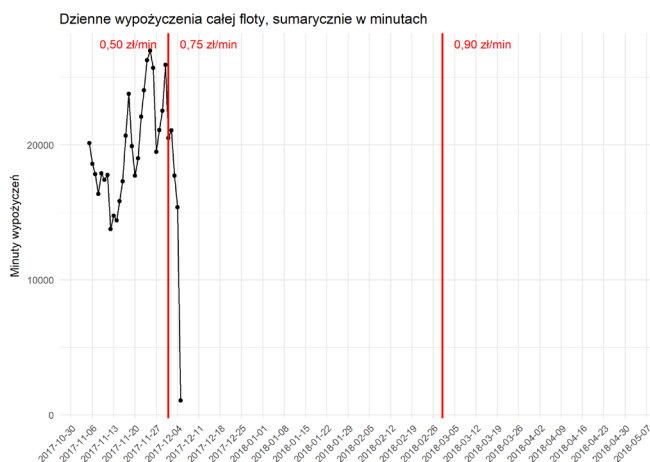
Zgodnie z zapowiedziami właściciela wypożyczalni, docelowa liczba 190 Nissanów Leaf została osiągnięta pod koniec stycznia 2018, w lutym do sieci włączono także dziesięć dostawczych e-NV200.

Ile samochodów dostępnych było do wypożyczenia każdego dnia?



Do połowy stycznia klientom musiało wystarczać 120-130 samochodów dostępnych każdego dnia, dopiero po końcowej dostawie Leafów liczba dostępnych aut przekroczyła półtorej setki.

Przez ile minut dziennie wypożyczono łącznie wszystkie auta?



Coś się nie zgadza: od 5 grudnia nie mamy żadnych danych o wypożyczeniach. Szybkie spojrzenie w logi pozwala stwierdzić, że tego dnia o godzinie 7 rano na serwery Vozilli wgrano aktualizację (usługa przez kilka minut leżała), po której z publikowanych danych znikły informacje o pojazdach wypożyczonych (status RENTED) oraz serwisowanych (UNAVAILABLE).

Oczywiście nie pogodzimy się tak łatwo z niekompletnością danych, jednak najpierw sprawdzimy, co uda nam się wyciągnąć z zapisu pierwszego miesiąca.

Stacje ładowania, trasy przejazdu, itd.

Informacje przedstawione w tym rozdziale zostały wyciągnięte z danych listopadowych, wtedy API Vozilli zwracało informację o pozycji wszystkich pojazdów, w tym wypożyczonych i serwisowanych.

Wielu klientom Vozilli przemknęło przez głowę pytanie:

Gdzie ładuje się samochody?

W końcu przy dedykowanych parkingach nie ma stacji ładowania, zaś w samochodach brak kabli. Firma zapowiada budowę kilkunastu stacji szybkiego ładowania w całym mieście, jednak na początku samochody ładowano w trzech miejscach.

Lokalizacja miejsc ładowania samochodów Vozilli

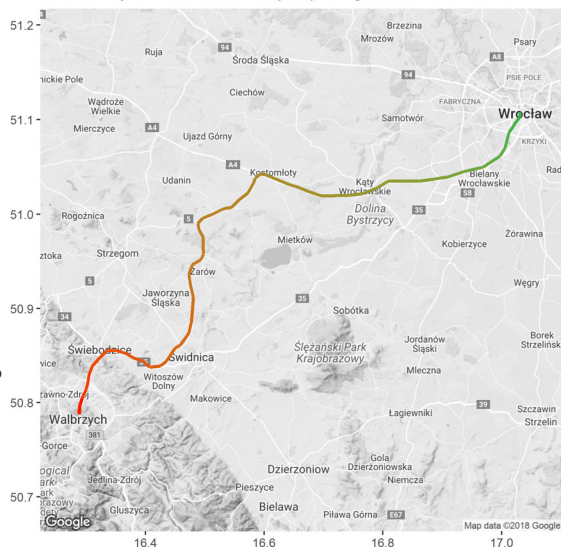


Widzimy, że główną bazą jest siedziba Vozilli przy ul. Kochanowskiego 30, z lokalizacjami zapasowymi pod Narodowym Forum Muzyki i przy ul. Wilanowskiej (Osiedle Sobieskiego). Jak zorganizować logistykę zwożenia aut do naładowania? Zdrowy rozsądek podpowiada, że po rozładowany samochód najłatwiej dojechać..

naładowanym samochodem! Obliczenia wydają się potwierdzać hipotezę - **niemal połowa naładowanych samochodów jest wprowadzana do wypożyczenia nie dalej niż 50 metrów od innego auta wyłączanego z obiegu.**

Spójrzmy na zarejestrowane przejazdy. **Komu udało się pokonać największy dystans od startu?** Okazuje się, że już pierwszej nocy ktoś pojechał „na dzidę” z Wrocławia do... Wałbrzycha. Na dystansie przeszło 80 km udało mu się zejść do 6% baterii. Auto wróciło do Wrocławia tego samego dnia po południu, prawdopodobnie na lawecie.

Jedna z wycieczek zakończonych poza granicami Wrocławia

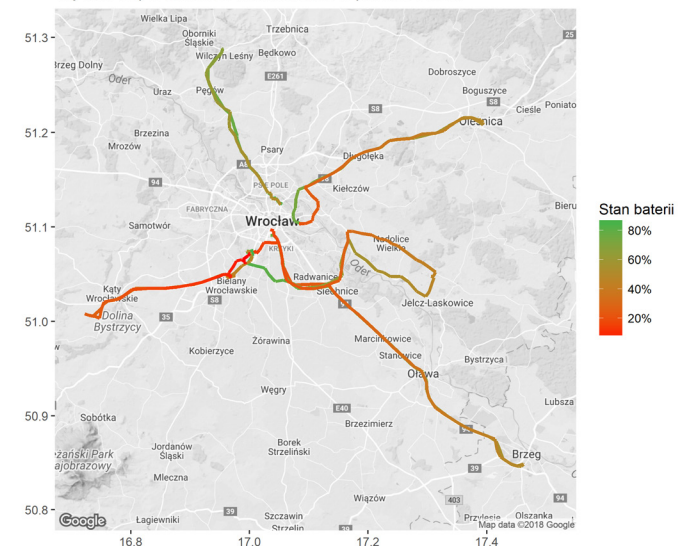


Kto pojechał na najdłuższą wycieczkę, z której wrócił dokładnie w miejsce startu?



Rekordzista wybrał się z Wrocławia do Brzegu - ponad 40 km w jedną stronę, czas podróży 147 minut, zużycie 73% baterii. Były też wycieczki do Oleśnicy, Kątów Wrocławskich, Jelcz-Laskowice czy Obornik Śląskich.

Wycieczki poza miasto zakończone w pobliżu startu

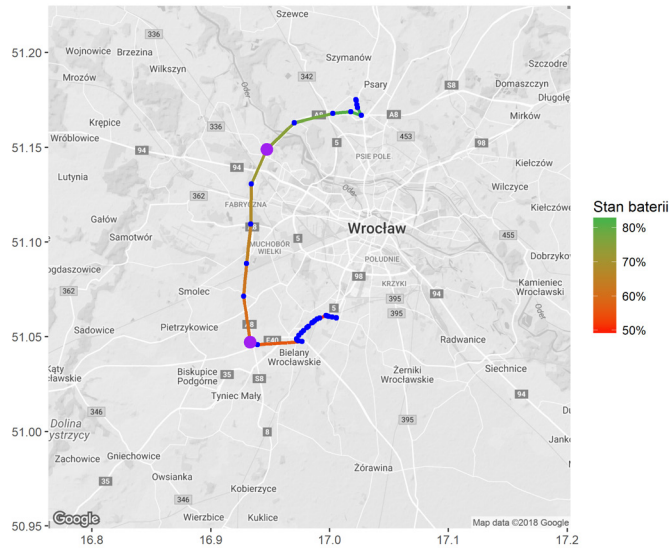


Dozwolona prędkość była przekraczana przez niejednego użytkownika Vozilli.

W przebytych trasach poszukałem największej odległości, jaką samochód przebył w ciągu 5 minut - w większości przypadków pobijanie rekordów prędkości miało miejsce na Autostradowej Obwodnicy Wrocławia.

Poniżej przykład trasy, w której 12 km od okolic Stadionu Miejskiego do węzła Wrocław-Południe udało się pokonać w 5 minut (144 km/h). Większe prędkości osiągnęto na krótszych dystansach, ale w takim przypadku do wiarygodnej analizy potrzebowalibyśmy zapisu pozycji o większej częstotliwości lub gwarantowanym interwale.

Przykład szybkiej jazdy na Autostradowej Obwodnicy Wrocławia



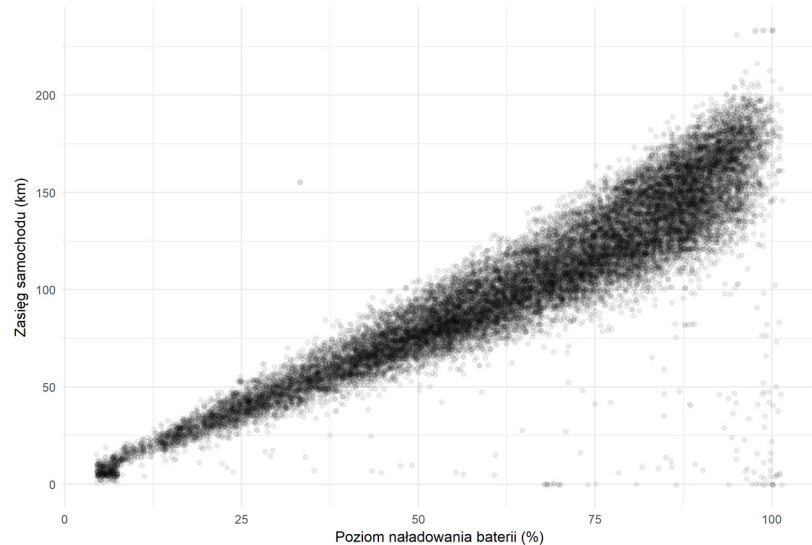
+ + +
 + + +
 + + +

Według danych producenta, Nissan Leaf z baterią 30kWh przejedzie od 100 km (zimą po mieście w korku, włączone ogrzewanie) do 200 km (przy wiosennej jeździe rekreacyjnej, brak ogrzewania, brak klimy, 60 km/h). Stan naładowania baterii oraz szacowany zasięg są prezentowane w samochodzie na wyświetlaczu oraz w serwisie online i aplikacji Vozilli. Przyjrzyjmy się zatem zależności między tymi dwiema wartościami.

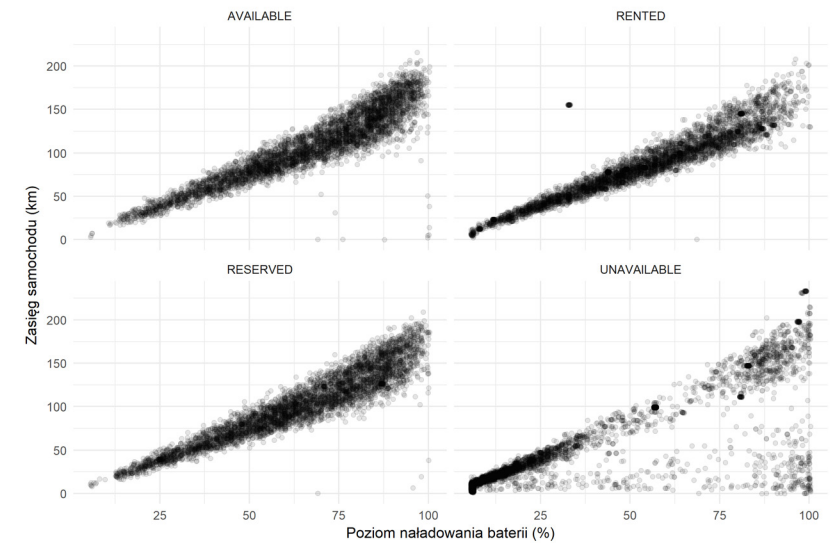
Widzimy sporo zaburzonych odczytów, w których pełna bateria daje zasięg bliski zeru. Po rozbiciu na statusy okazuje się, że to głównie status niedostępny sieje takimi zakłóceniami.

Zasięg samochodów elektrycznych.

Prognozowany zasięg w zależności od naładowania baterii

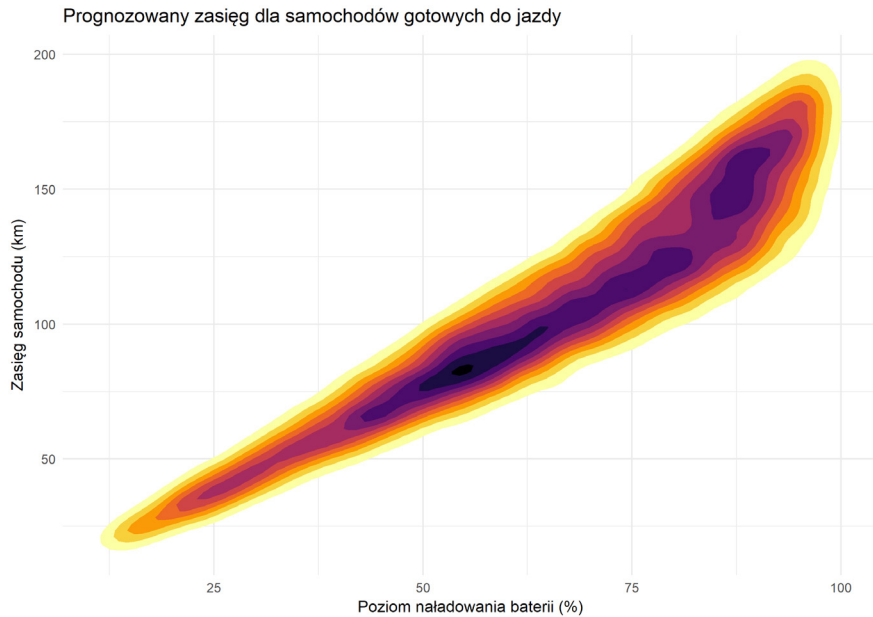


Prognozowany zasięg w zależności od naładowania baterii i statusu



Tu małe wyjaśnienie - istnieje jeszcze status „RETURNED”, ale w całym tekście go pomijamy; w pierwszym miesiącu wystąpił nieco ponad sto razy i nie bardzo da się go powiązać z akcjami użytkowników końcowych.

Rezygnujemy z samochodów niedostępnych. Punktów nadal jest za dużo, spróbujmy większej próbki i mapy gęstości 2D.



Jest znacznie lepiej. Wyraźnie widzimy „główny ciąg” z praktycznie liniową zależnością: w pełni naładowany Leaf przejedzie około 150 km.

Wpływ temperatury na zasięg.

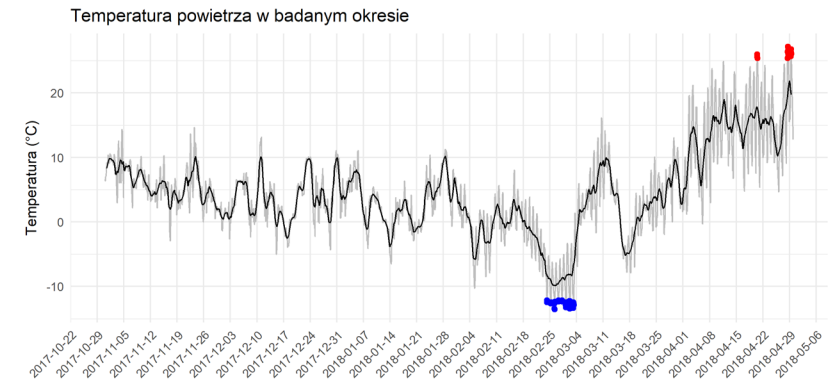
Ciekawe, jak duży jest wpływ temperatury otoczenia na zasięg samochodów Vozilli.

Auta elektryczne są szczególnie wrażliwe na temperatury poniżej zera

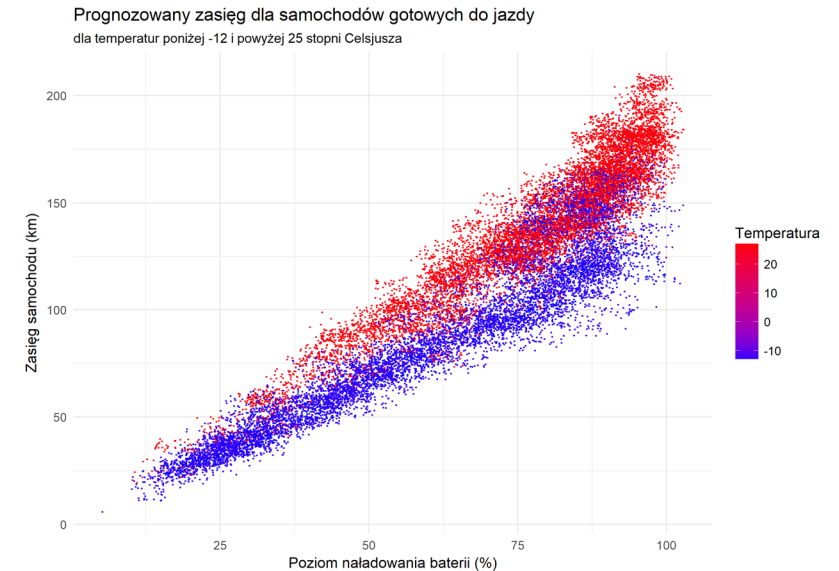
- spada pojemność ogniw, część energii jest zużywana na ich podgrzewanie, sporo zjada też ogrzewanie kabiny, kierownicy i foteli. Krótkie dystanse sprzyjają ustawieniu nagrzewnicy na maksimum.



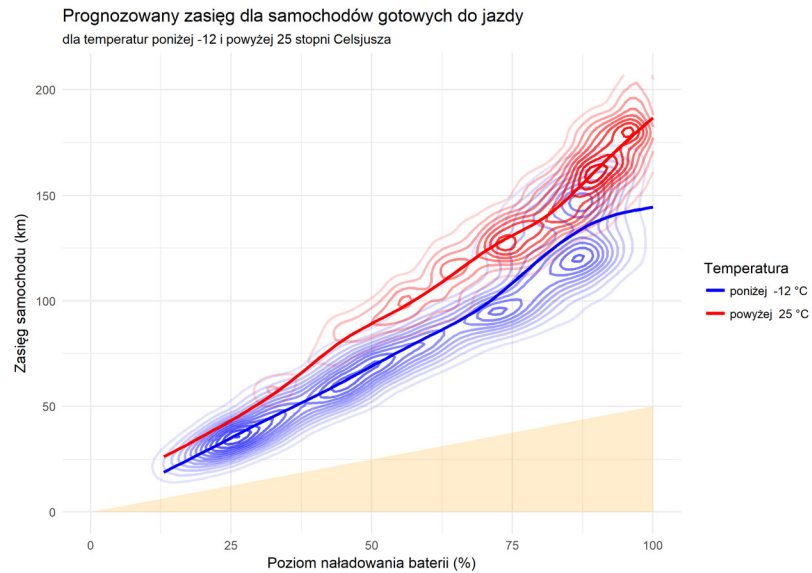
Musimy zacząć od sięgnięcia do danych Instytutu Meteorologii i Gospodarki Wodnej, gdzie odnajdujemy pliki CSV z archiwalnymi pomiarami stacji meteo rozsianych po całym kraju. Duże różnice temperatur między lutym a kwietniem 2017 pozwoliły porównać zachowanie pojazdów w temperaturach niższych od -12 i wyższych niż 25 stopni Celsjusza.



Wzmy więc tyle samo próbek z górnego i dolnego przedziału temperatur i umieścmy je na jednym wykresie:



Na wizualizacji dostrzegamy, że w „głównym ciągu” kolor czerwony sytuuje się delikatnie wyżej. Widzimy jednak, że ewidentnie błędne prognozy zasięgu (szacowany zerowy dystans) zakłóca analizę, więc w następnym kroku odetniemy je.



↑
Na wykresie widać wyraźne różnice, nie możemy jednak być pewni, że za niższe prognozy odległości na pewno odpowiada niższa temperatura otoczenia. Dlaczego?

Najwyższe temperatury odnotowano w środku dnia, a najniższe w nocy - być może samochód nieużywany od wielu godzin zaniża szacunki, biorąc pod uwagę samorozładowanie baterii. Ekstremalne temperatury dzieliły dwa miesiące - może nowe samochody włączane do floty prognozują inaczej, niż po kilku(nastu) tygodniach służby? **Aby wyeliminować wpływ takich zmiennych, potrzebowalibyśmy danych z dłuższego okresu, najlepiej kilku sezonów.**

Rekonstrukcja brakujących danych.

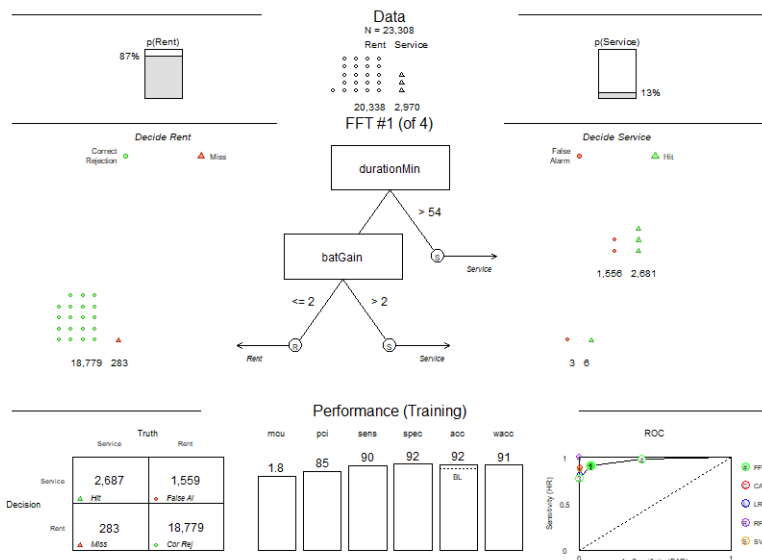
Jak już wspomniano, od 5. grudnia gromadzone dane obejmują wyłącznie samochody dostępne i zarezerwowane. Gdy auto jest wypożyczone lub serwisowane, znika z API. Do dalszej analizy potrzebujemy opracować metodę rekonstrukcji brakujących danych. Oczywiście nie odtworzymy lokalizacji samochodu, ale możemy próbować odgadnąć, co się z nim działo.

+ + + W pierwszym kroku minutowe obserwacje całej floty przekształcamy w obserwacje zdarzeń każdego pojazdu z osobna. Przykład: 600 obserwacji minutowych pojedynczego samochodu może przekładać się na cztery następujące po sobie zdarzenia: oczekiwanie (200 minut), rezerwacja (15 minut), wypożyczenie (45 minut) i ponowne oczekiwanie (300 minut). W praktyce 30,5 mln obserwacji przełożyło się na 476 tysięcy zdarzeń.

Następnie model danych opisujący pojedyncze zdarzenie wzbogacamy o zmienne wyliczane, np. czas trwania czy zmianę naładowania akumulatora w punktach procentowych. Dostępne są też dane opisujące pozycję i stan pojazdu na początku i końcu trwania zdarzenia.

W przedostatnim kroku do listy zdarzeń odczytanych z danych źródłowych dodaliśmy pozycje zrekonstruowane, odpowiadające „dziurom” w danych obserwacyjnych. Nieznany pozostał jedynie typ (status) tych zdarzeń.

Do rekonstrukcji typu zdarzenia użyta została heurystyka zbudowana za pomocą algorytmu Fast-and-Frugal Decision Trees z pakietu FFTrees. Zestawem danych uczących był listopadowy zbiór zapisów o wypożyczeniach i serwisowaniu całej floty. Wynikiem działania algorytmu jest drzewo decyzyjne, w opisywanym przypadku wyglądało ono następująco:



Wyznaczony algorytm jest prosty - mamy do czynienia z serwisowaniem, gdy zdarzenie trwa dłużej niż 54 minuty lub stan baterii rośnie o więcej niż dwa punkty procentowe. Osiągnięta dokładność (accuracy) tej heurystyki to 92% i mogłaby wydawać się relatywnie niska, gdyby nie obserwowane zaburzenia w danych źródłowych - odnajdujemy tam np. kilkudobowe wypożyczenia albo tygodniowe serwisowanie bez ładowania. Niestety, jedynie operator usługi wie o takich wydarzeniach jak stłuczki i naprawy, informacji o nich nie znajdziemy w API.

Jak trafnie zrekonstruowaliśmy dane o wypożyczeniach? Wiemy, że poszło całkiem nieźle - na początku marca 2018 przedstawiciele Vozilli opublikowali [statystyki opisujące pierwsze cztery miesiące działania wypożyczalni](#):

- 80 tysięcy wypożyczeń (w naszych obliczeniach - 73 tys.),
- prawie 1,3 miliona wypożyczonych minut (u nas - 1,5 miliona, ale po odjęciu nienaturalnie długich „wypożyczeń” z pierwszego miesiąca trafiamy w punkt).
- średni czas wypożyczenia to wg Vozilli 16 minut czyli tyle samo, ile mediana w naszych kalkulacjach.

Potwierdziłmy w ten sposób, że



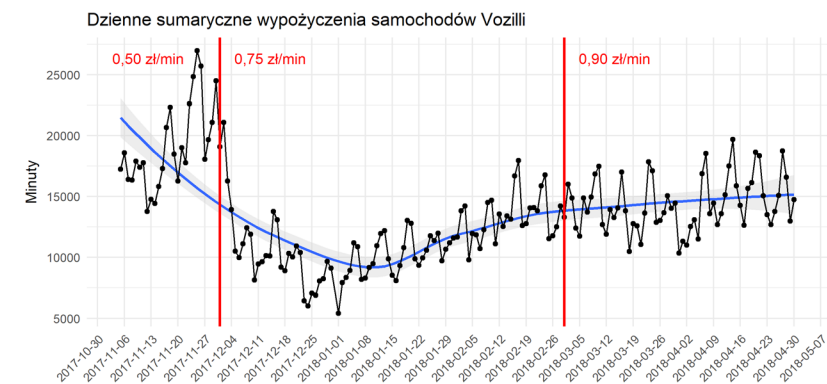
przyjęta metodologia sprawdziła się



a uzyskane wyniki nie powinny odbiegać od wartości rzeczywistych o więcej niż 10%.

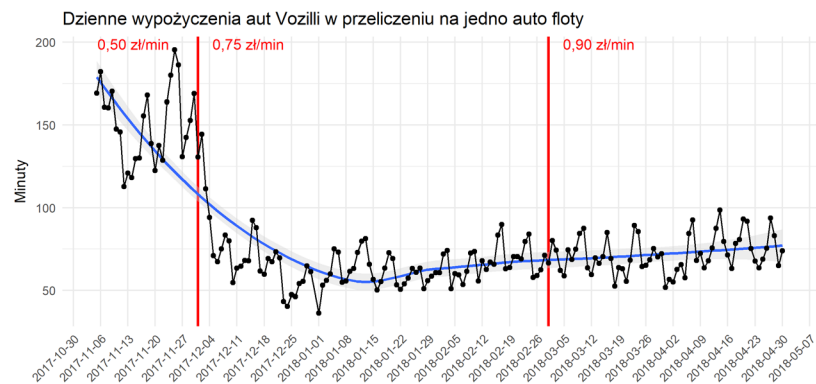
Statystyki z odtworzonych danych.

Po zastosowaniu opisanego wyżej algorytmu jesteśmy w stanie oszacować liczbę wypożyczeń w okresie od listopada 2017 do końca kwietnia 2018. Dzienna liczba wypożyczeń (liczona w minutach) przedstawia się następująco:

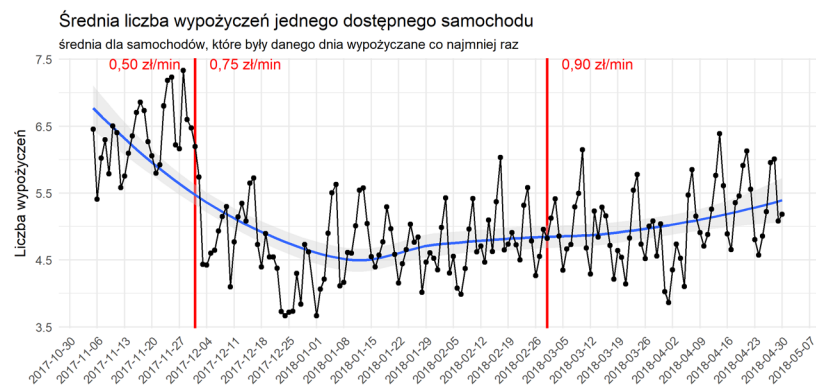


Przez pierwszy miesiąc wypożyczenie auta kosztowało 50 groszy za minutę, rekordowy wynik zrobiła masa chętnych na przejechanie się samochodem elektrycznym. **W grudniu cena wzrosła do 75 groszy za minutę, a ruch spadł o połowę.** Od początku roku 2018 liczba wypożyczeń rośnie, choć w marcu cena została podniesiona do 90 groszy za minutę.

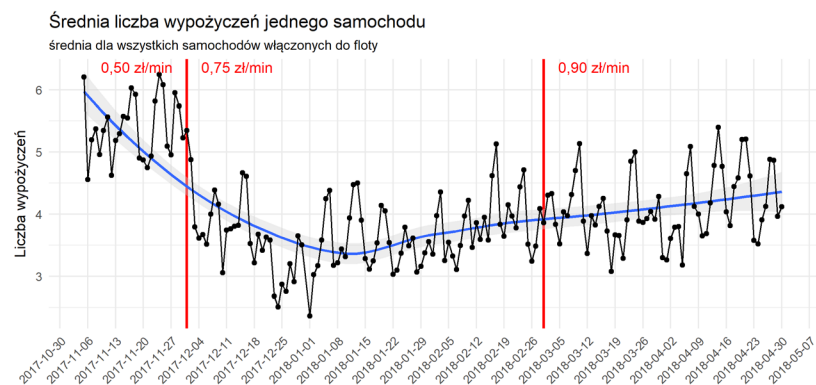
Wypożyczenia w minutach na jeden samochód floty:



Liczba wypożyczeń w przeliczeniu na jeden samochód dostępny danego dnia:



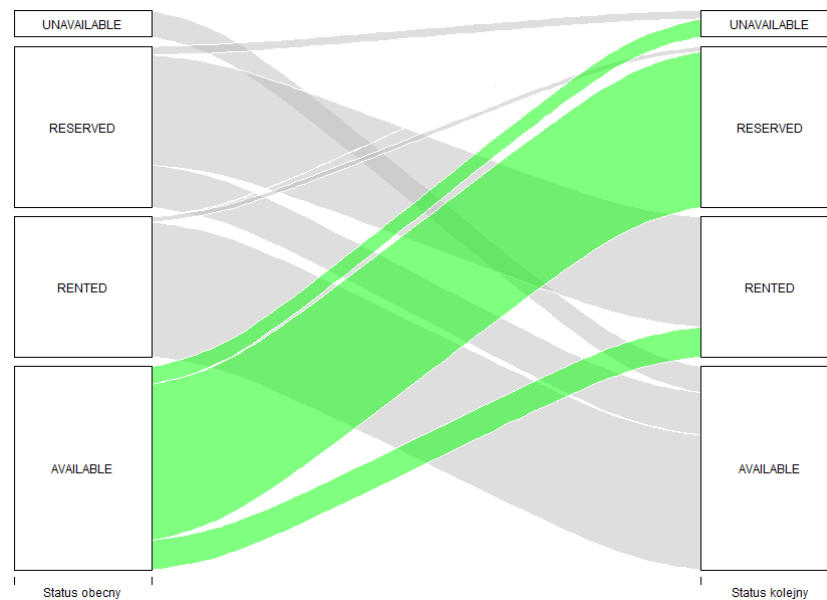
Liczba wypożyczeń w przeliczeniu na liczbę samochodów floty:



Inne statystyki.

Jak często rezerwujemy auto przed wypożyczeniem?

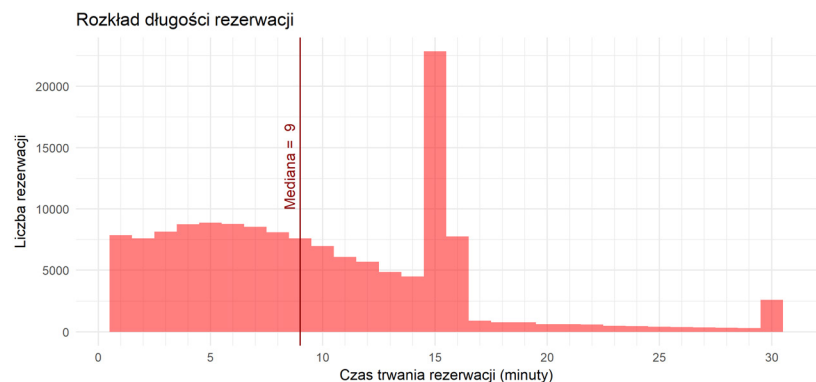
W przeważającej części przypadków. Około trzech czwartych rezerwacji kończy się wypożyczeniem.



Ile czasu trwa typowa rezerwacja?

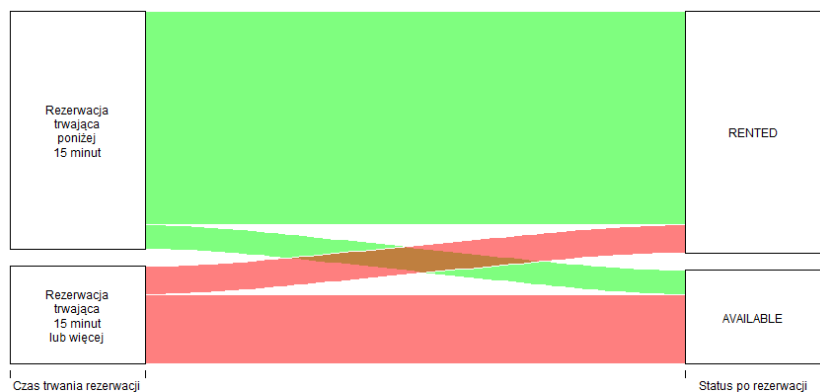
Mediana to dziewięć minut, ale dominują piętnastominutowe rezerwacje „przytrzymujące” auto w jednym miejscu.

Rezerwacje dłuższe niż kwadrans nie powinny być możliwe, być może na wykresie poniżej widzimy efekt szybkiego zwolnienia i ponownej rezerwacji auta (wartości równe lub wyższe niż 30 minut zostały na wykresie zsumowane).

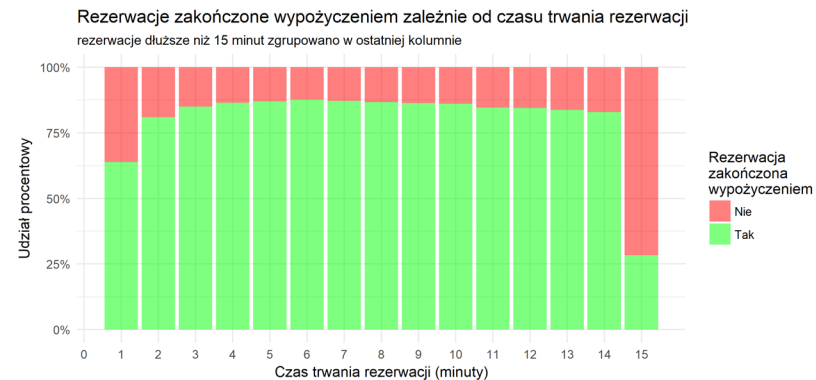


Czy rezerwacja trwająca maksymalny okres kończy się wypożyczeniem? Otóż nie

większość długich rezerwacji zostaje zwolniona

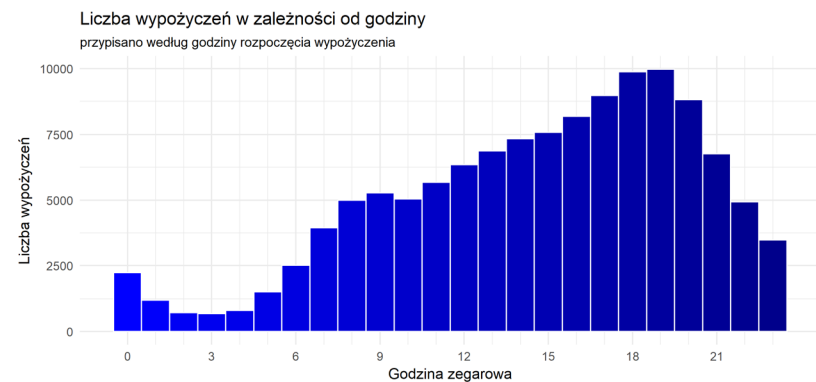


Ile rezerwacji o danej długości kończy się anulowaniem? W przypadku jedynominutowek przyczyną mogą być problemy z aplikacją, której zdarzało się zarezerwować samochód inny od oczekiwanego. Najdłuższe wypożyczenia (ostatnia kolumna to rezerwacje kwadransowe i dłuższe) to w 2/3 przypadków przytrzymywanie auta „na potem”.



O której godzinie najczęściej wypożycza się auta?

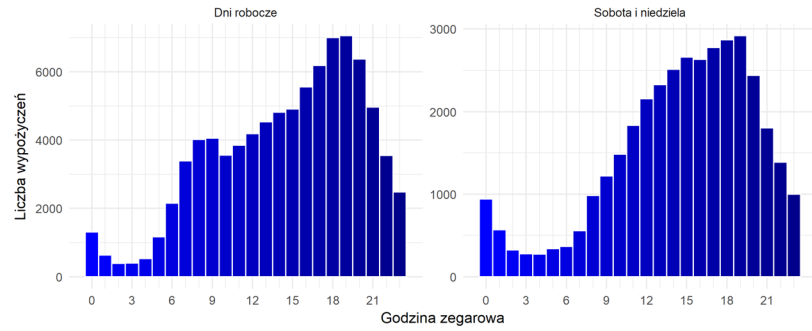
Między 17:00 a 20:00.



Czy weekendowe wypożyczenia jakoś się różnią?

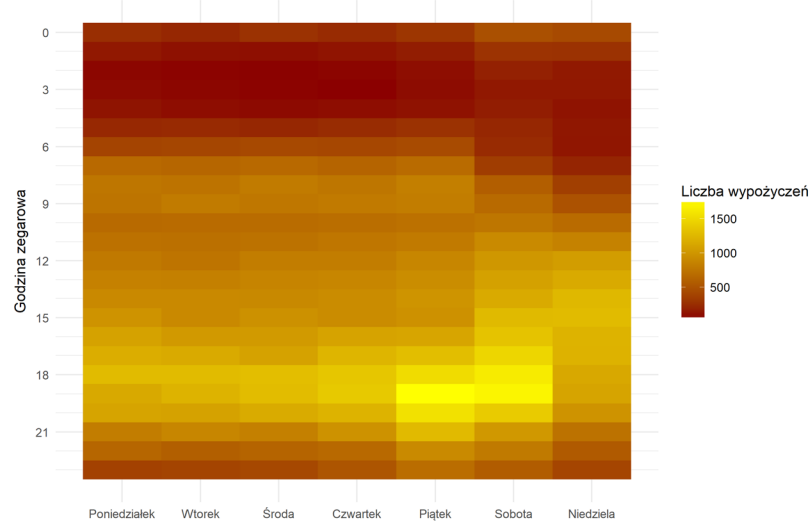
Nieznacznie, w weekend znika zafalowanie około godziny 8, za to godziny szczytu zaczynają się już o 15.

Liczba wypożyczeń wg godzin w dni robocze i weekendy
przypisano według godziny rozpoczęcia wypożyczenia



Wypożyczenia z podziałem na dni tygodnia: piątkowy wieczór świeci jak żarówka, w sobotę i niedzielę śpimy dwie godziny dłużej.

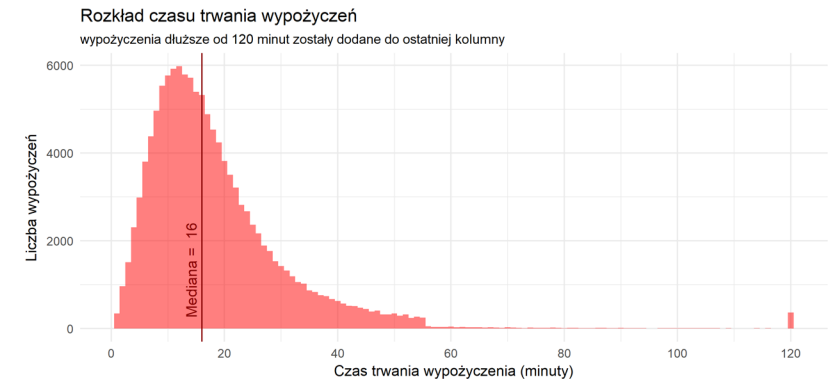
Liczba wypożyczeń z podziałem na godziny i dni tygodnia
przypisano według godziny rozpoczęcia wypożyczenia



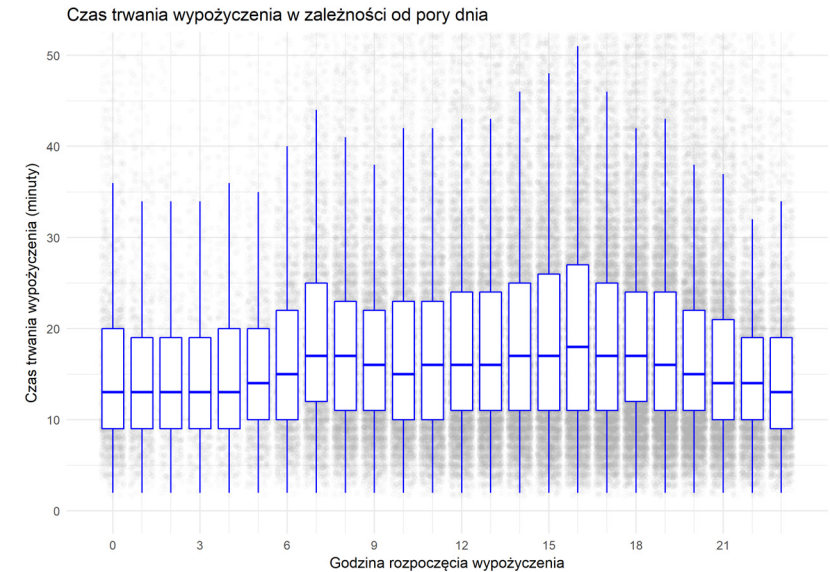
Ile czasu trwa typowe wypożyczenie?

Mediana to 16 minut, średnia wynosi 19 minut

(zdroworozsądkowo odfiltrowano wypożyczenia dłuższe niż 10 godzin; wypożyczenia trwające między 2 a 10 h zostały na wykresie zsumowane).



Czy czas wypożyczenia zależy od pory dnia?



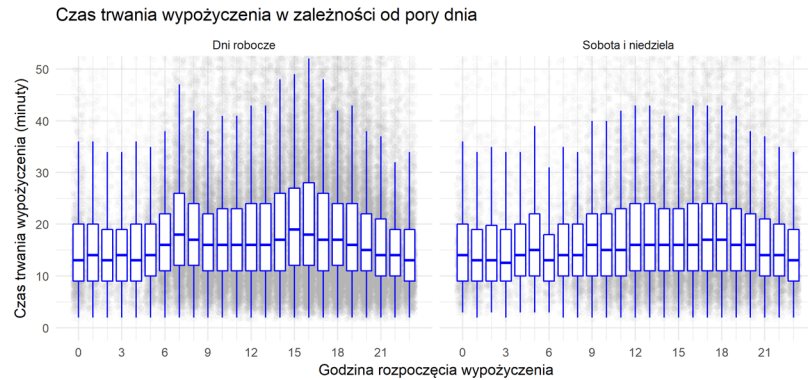
To samo z podziałem na dni tygodnia



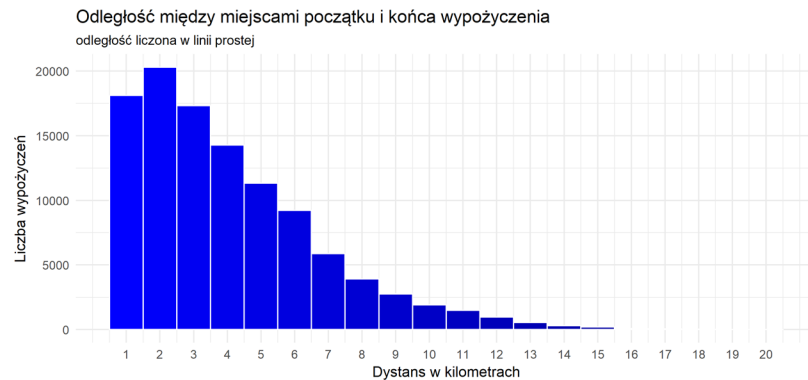
w weekendy jeździmy w godzinach szczytu nieco krócej



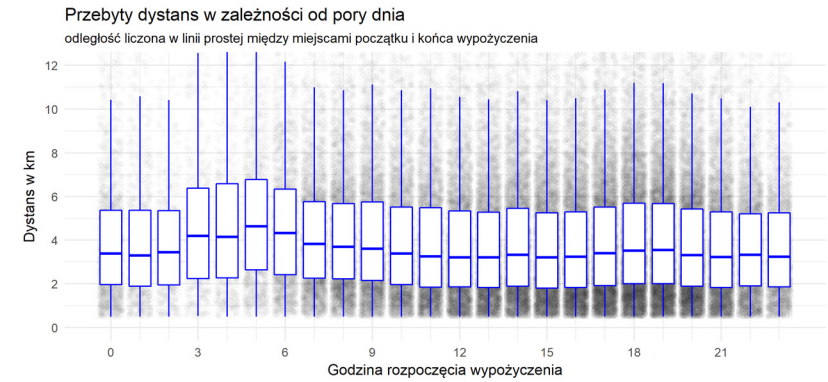
Odpowiadają za to krótsze dystanse czy mniejszy ruch na drogach?



Odległość w linii prostej od miejsca startu do miejsca zakończenia wypożyczenia. **Vozilla podaje, że średni pokonywany dystans to 7 km**, nasz wynik jest niższy przez kierowców robiących pętlę i wracających w pobliże startu, wpływ ma też układ komunikacyjny miasta przeciętego siecią rzek i kanałów.

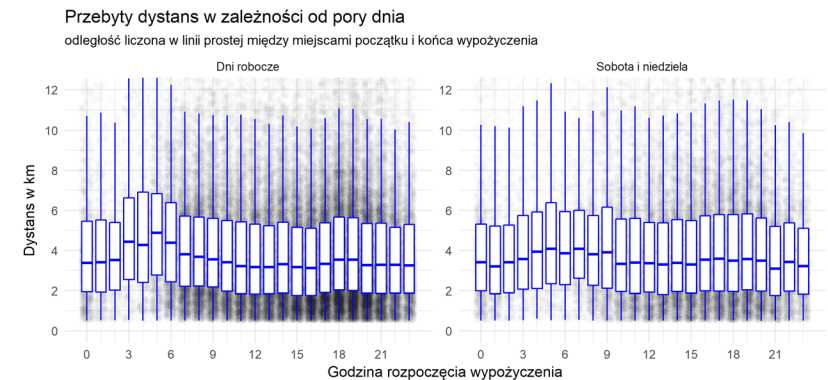


Dystans w zależności od pory dnia?

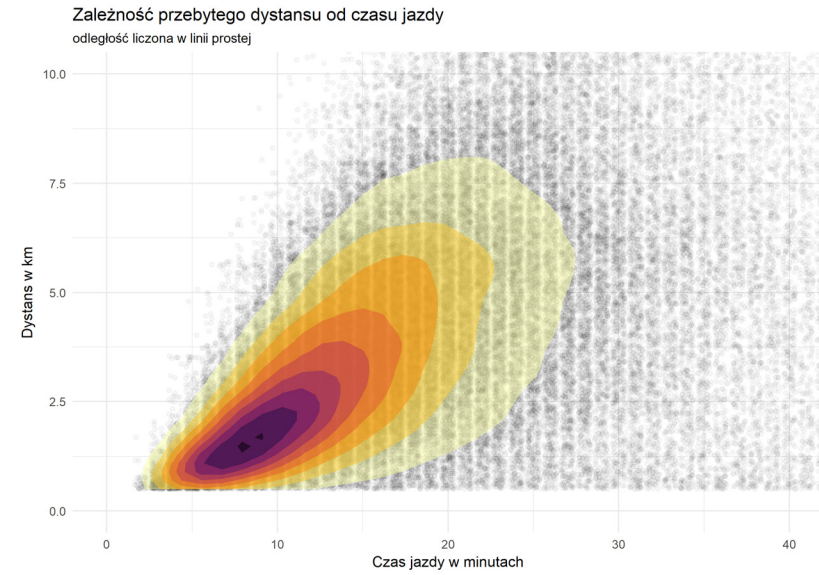


To, co powyżej, z podziałem na dni robocze i weekendy.

Widzimy, że rozkład odległości pokonywanych w weekendowe popołudnia nie różni się od dni roboczych, stąd wniosek, że za krótszy czas jazdy w weekend odpowiada mniejszy ruch uliczny

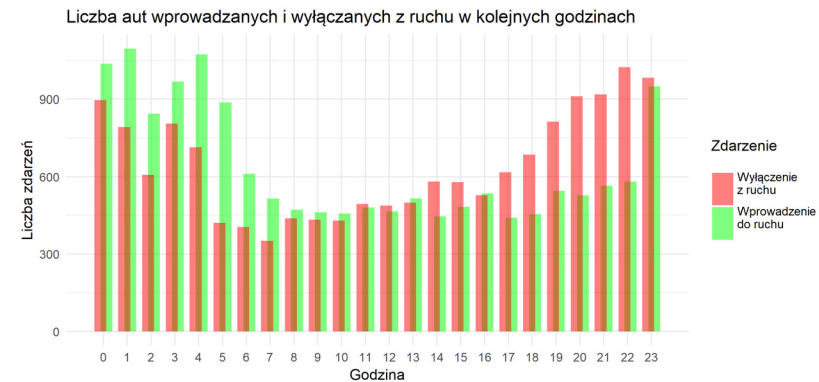


Dystans w zależności od czasu trwania podróży?



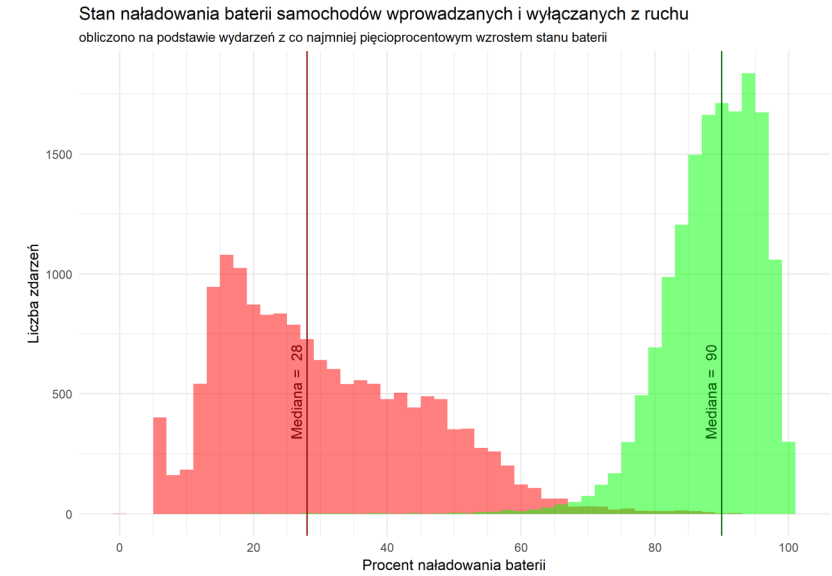
Popołudniowy i wieczorny popyt skutkuje stopniowym wycofywaniem aut z ruchu, kolejka do ładowania obsługiwana jest aż do rana.

Przedpołudnie to czas równowagi.

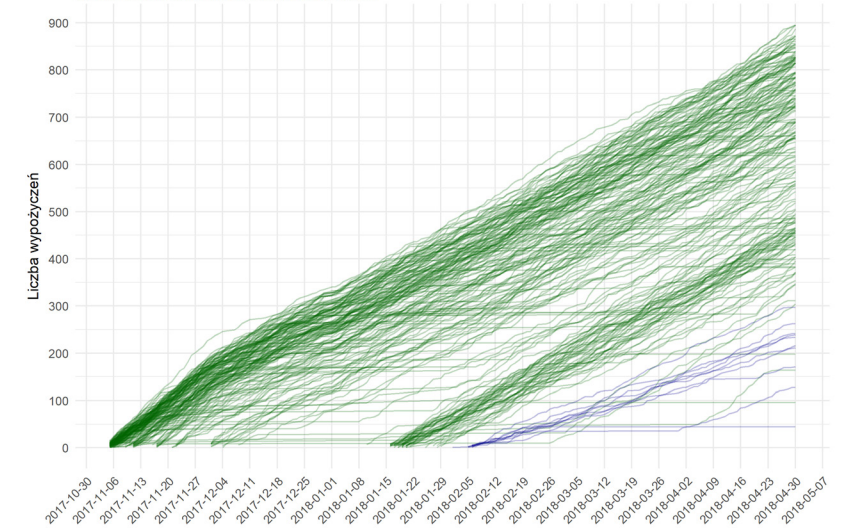


Przy jakim stanie baterii samochody są ściągane z trasy, a przy jakim wracają do wypożyczenia?

Mediana to odpowiednio 28% i 90%.

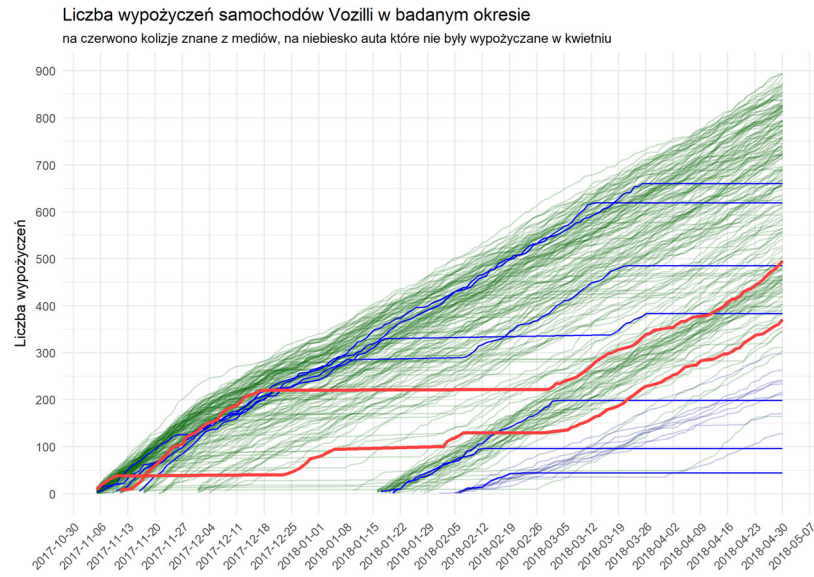


Liczba wypożyczeń samochodów Vozilli w badanym okresie kolorem niebieskim oznaczono dostawce e-NV200

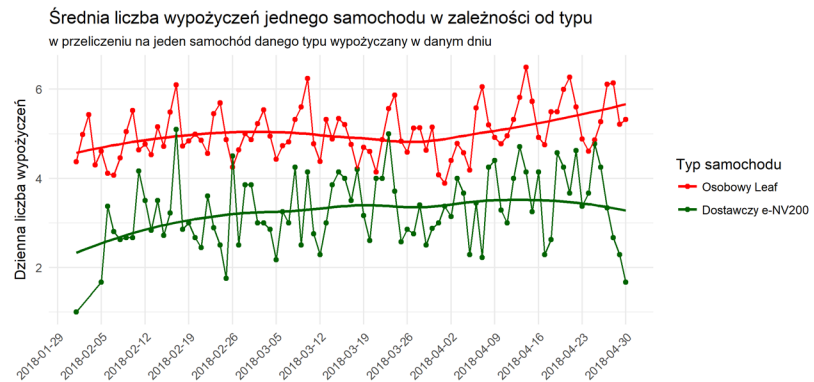


Co z opisywanymi w gazetach słuzkami aut Vozilli z 10 listopada i 18 grudnia?

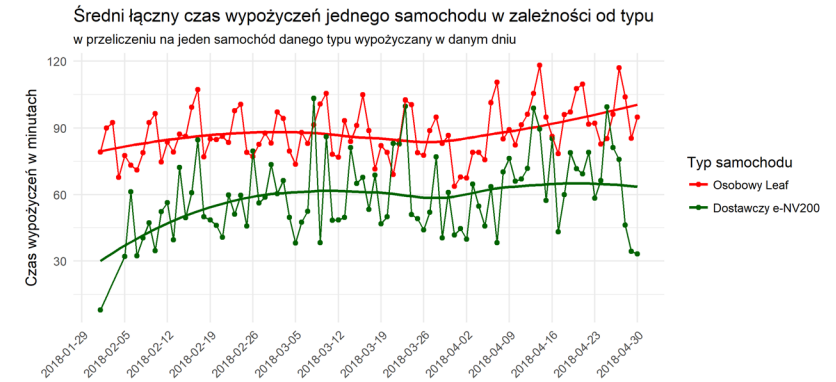
Poniżej zaznaczone są na czerwono, dodatkowo na niebiesko te auta, które nie były wypożyczone w kwietniu.



Na zakończenie zobaczymy, jak radzą sobie dostawcze Nissany e-NV200, które zadebiutowały w sieci Vozilla z początkiem lutego 2018. Widać, że w liczbie wypożyczeń na jedno auto danego typu jeżdżące danego dnia - odstają wyraźnie od osobowych Leafów.



Może więc dostawczaki nadrabiają dłuższymi wypożyczeniami? Również nie, ich średnia użyczenie jest niższa niż osobowych.

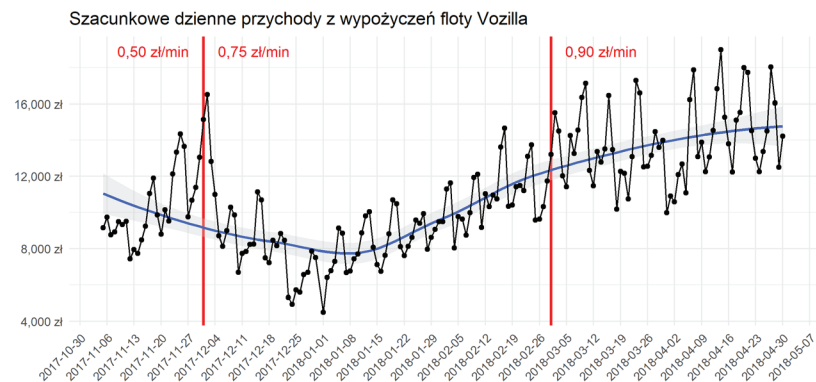


Jakie przychody notuje Vozilla?

Poniższe szacunki należy traktować ostrożnie - mogą być obciążone błędem. Trzy główne powody:

- nasze próby odróżnienia wypożyczeń od serwisowania nie zawsze się udają (zaniżamy liczbę wypożyczeń)
- kierowca ma kilka darmowych minut na przygotowanie auta do jazdy (tu zawyżamy przychody)
- opłata podczas rezerwacji i postoju wynosi jedynie 10 groszy za minutę; nie wiadomo, jaką część dłuższych wypożyczeń stanowi postój (tu też zawyżamy przychody)

W pierwszym miesiącu rekordowa liczba chętnych wypróbowała samochód elektryczny Vozilli, bez wątpliwości wpływ na to miała cena obniżona o połowę względem zapowiadanej stawki regularnej. Dzielne przychody brutto osiągnęły historyczne minimum w okolicach Bożego Narodzenia 2017, jednak od tamtego czasu systematycznie rosną:



W badanym okresie liczba aut w sieci wzrosła prawie dwukrotnie.

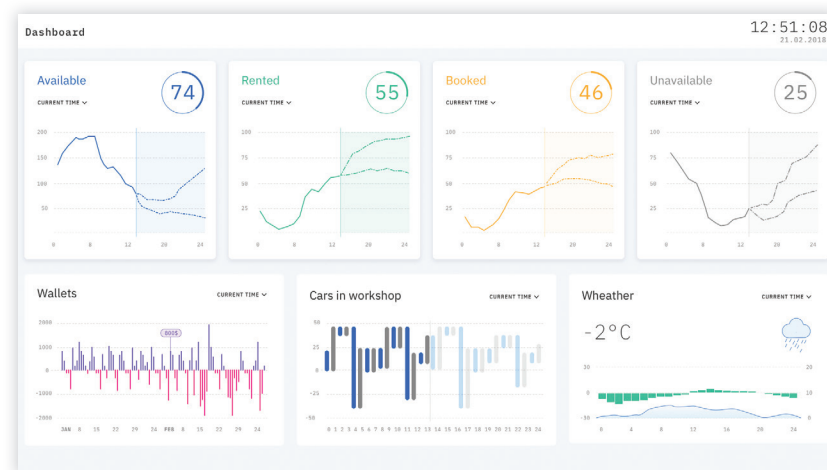
” Jak wyglądały przychody w przeliczeniu na jeden samochód floty?

Wg naszych szacunków - na przełomie lat 2017/18 spadły poniżej 50 zł na dobę. Od tamtego czasu średnia stabilnie rośnie, osiągając w ciągu czterech miesięcy poziom 75 zł dziennie.



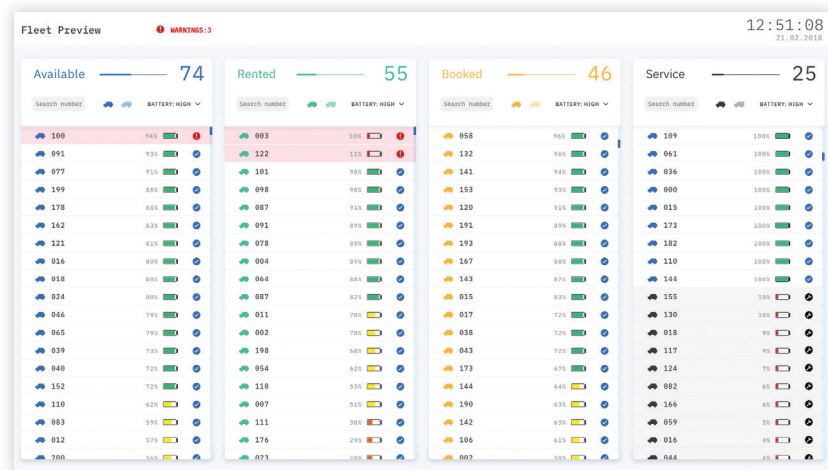
Dodatek 1 Dashboard

W systemach do zarządzania flotą pojazdów przydatnym narzędziem jest tzw. dashboard, czyli **graficzna wizualizacja najważniejszych parametrów opisujących bieżący stan sieci**. Na kilku dużych ekranach wiszących na ścianie można zwięźle przedstawić informacje, które pomogą dyspozytorowi w podejmowaniu trafnych decyzji. Poniższe makiety, przygotowane przez zespół designu PGS Software, mogłyby być pierwszym etapem prac nad taką wizualizacją.



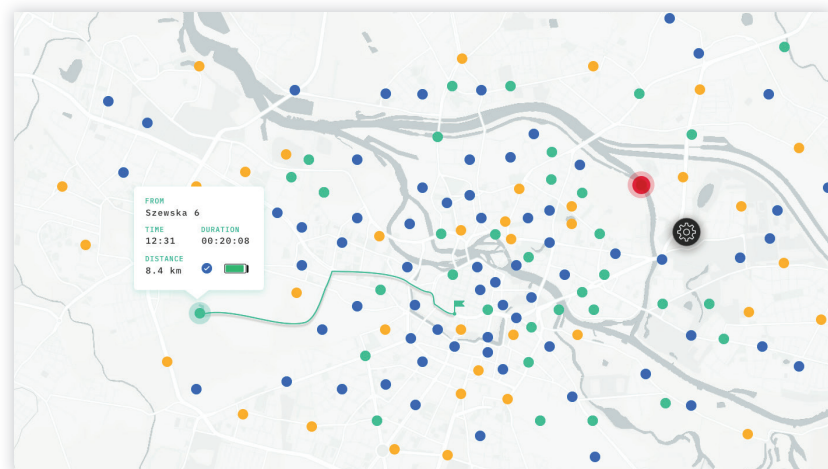
Podstawową informacją opisującą stan bieżący jest liczba samochodów dostępnych, zarezerwowanych, wypożyczonych oraz niedostępnych

(tu potencjalnie z podziałem na auta do naładowania oraz serwisowane)



Osobnym ekranem może być mapa miasta z reprezentacją pozycji samochodów

choć przy dwustu samochodach czytelne przedstawienie wszystkich naraz może być pewnym wyzwaniem.



Znacznie bardziej wartościowe może być jednak prognozowanie metryk z wielogodzinnym wyprzedzeniem - bazując na danych historycznych i parametrach bieżących. Oto, jakie informacje moglibyśmy przedstawić w formie wizualnej lub opisowej:

- Jaki procent planu dziennego lub tygodniowego uda się zrealizować przy utrzymaniu bieżącego tempa wypożyczeń
- Jaki jest maksymalny osiągalny pułap przychodów przy założeniu, że wszystkie stacje ładowania będą wykorzystane w stopniu maksymalnym
- Jak będzie spadek dostępności samochodów (a więc i przychodów), jeśli zmniejszy się liczba osobogodzin poświęconych na ich obsługę (zwożenie aut do ładowania i rozwożenie naładowanych)
- Jak uszeregować auta do zwózki i przestawienia, aby zmaksymalizować użycie floty (z opcjonalnym rozwożeniem wielu kierowców naraz, uwzględnieniem czasu przejazdu po mieście, zajętości stanowisk ładowania, zmiennych kosztów pracy, zmienności popytu itd.)
- Do jakiego poziomu ładować aktualnie podłączone samochody, aby zmaksymalizować użycie floty zgodnie z prognozowanym popytem?

Algorytmy optymalizacyjne mogłyby na bieżąco korygować powyższe prognozy i zalecenia. Zakres danych wejściowych można wzbogacić np. o **lokalizację popularnych wydarzeń sportowych lub kulturalnych** - podstawienie w takim dniu kilkudziesięciu samochodów pod Stadion Miejski (oddalony o 6 km od centrum) gwarantowałoby ich pełne wykorzystanie na trasach dłuższych od przeciętnej.



Jak tanio i masowo przemieszczać auta Vozilli?

Jedną z odpowiedzi jest odwrotna aukcja - zaoferowanie użytkownikom zniżki na przejazd samochodem zaparkowanym na odludziu. Jeśli wiemy z doświadczenia, że naładowane auto spędzi tam ponad X godzin, a utracone korzyści z jazdy po centrum miasta to Y zł, operatorowi może opłacać się wręcz zapłacić użytkownikom końcowym za przejazd autem do wyznaczonej strefy wysokiego popytu.

Dodatek 2 Przetwarzanie w chmurze

Pierwsza wersja skryptów przetwarzających dane była bardzo wolna - import i wstępne przekształcenie czteromiesięcznych danych trwało na typowym komputerze (CPU i5-4570, 3.20 GHz, 4 rdzenie, 16 GB RAM, SSD) przeszło 14 godzin. Gotowy zestaw wyników można było oczywiście zapisywać i odtwarzać na żywo, jednak iteracyjne ulepszanie i testowanie procedur było mocno utrudnione.

Wersja druga korzystała z wielowątkowości - dzięki rozłożeniu zadań na kilka jednostek przetwarzających procesora, czas przetwarzania skrócił się przeszło dwukrotnie, jednak nadal był nieakceptowalnie długi.

W dzisiejszych czasach najpopularniejszym podejściem do takiego problemu jest przeniesienie zadania na chmurę obliczeniową Amazon AWS, Microsoft Azure lub Google Cloud Platform.

Wśród załogi PGS Software największą popularność zdobył Amazon Web Services, w chwili pisania tych słów zgromadziliśmy już przeszło pół tysiąca akredytacji i ponad setkę certyfikatów AWS. W tu opisywanym przypadku

zdecydowałem się jednak na użycie Azure - Microsoft od dawna inwestuje w język R, a w swojej chmurze udostępnia gotowe usługi integracyjne dla tego środowiska.

Testy przeprowadziłem przy użyciu pakietu doAzureParallel służącego do rozpraszania obliczeń w klastrze maszyn wirtualnych działających w Azure Batch (jest to środowisko do wsadowego przetwarzania zadań, z takimi udogodnieniami jak np. automatyczne skalowanie zależne od rozmiaru danych wejściowych). Zestawienie klastra obliczeniowego sprowadza się tam do kilkunastu kliknięć i edycji jednego pliku konfiguracyjnego.

Próba rozłożenia zadania na kilkanaście lub kilkadziesiąt komputerów w chmurze nie przyniosła jednak oczekiwanych zysków, bo najpierw musimy przesać tam kilkanaście gigabajtów danych a potem zadbać, by odpowiednie pliki zostały rozpakowane na odpowiednich maszynach należących do klastra. Nie jest to specjalnie trudne, jednak wykracza poza możliwości pakietu do AzureParallel, tracimy więc elegancję rozwiązania opartego wyłącznie o platformę R. Przede wszystkim jednak - transfer i dekompresja takiej liczby plików sprawi, że pojedyncze uruchomienie zadania w chmurze nie przynosi realnych oszczędności czasowych. Z zupełnie inną sytuacją mielibyśmy do czynienia, gdyby klastr był używany do obliczeń cyklicznych - **wówczas możemy spodziewać się przyspieszenia proporcjonalnego do liczby maszyn.**

Skoro chmura nie pomogła, sięgnąłem po profiler i zacząłem optymalizować istniejący kod. Udało się to nadspodziewanie dobrze -



czas przetwarzania został skrócony przeszło dziesięciokrotnie.



Pierwszym krokiem w dobrą stronę było połączenie plików JSON tak, by zapis całego dnia znalazł się w jednym zbiorze. Parser języka R dużo lepiej radził sobie

z ładowaniem i parsowaniem jednego wielkiego pliku, niż tysiąca małych. Profiler wskazał też linię kodu odpowiadającą za 75% czasu późniejszego przetwarzania, tam wystarczyła drobna korekta algorytmu. Koniec końców wczytanie półrocznych danych trwa na pojedynczej maszynie nieco ponad kwadrans zaś ich późniejsza obróbka – około 2 godzin..

→ Cloud computing byłby przydatny do budowy i utrzymania dashboarda opisanego w Dodatku 1. Strumieniowe przetwarzanie informacji, szyna danych łącząca poszczególne komponenty, automatyczne skalowanie w zależności od chwilowych potrzeb - we wszystkich tych przypadkach AWS i Azure dostarczają gotowe klocki, na bazie których **możemy budować rozwiązania niezawodne, szybkie i efektywne kosztowo.**

– Dodatek 3

Przetwarzanie w chmurze

Jeśli chcesz przeprowadzić własne eksperymenty na danych Vozilli lub odtworzyć nasze obliczenia, odwiedź stronę <https://github.com/PGSSoft/vozilla>

Znajdziesz tam:

- archiwum zawierające 252 tysiące źródłowych plików JSON
- archiwum zawierające te same dane ale w formie 176 plików dobowych
- skrypty języka R służące do wczytywania, przetwarzania i wizualizacji danych
- workspace języka R zawierający wczytane dane

– O nas

PGS Software jest piątym w Polsce dostawcą oprogramowania na zamówienie. Obsługuje klientów na całym świecie, przede wszystkim z Europy Zachodniej. Spółka notowana jest na głównym parkiecie warszawskiej giełdy.

Jesteśmy partnerem dwóch największych dostawców rozwiązań chmurowych, dysponujemy certyfikatami AWS Advanced Consulting Partner i Microsoft Gold Certified Partner. Dostarczamy usługi dla rynków B2B i B2C: analizę biznesową, projektowanie UX i UI, programowanie i testy aplikacji internetowych oraz mobilnych. Specjalizujemy się w transformacji do chmury.

Mamy 3 centra rozwojowe w Polsce (Wrocław, Gdańsk i Rzeszów) oraz oddziały w Londynie, Monachium i Barcelonie.

www.pgs-soft.com
+48 71 798 2699
info@pgs-soft.com

PGS 
SOFTWARE