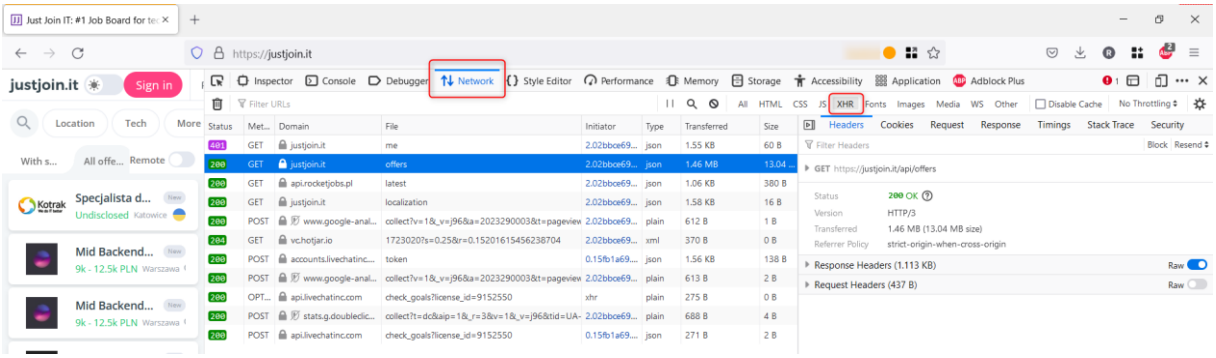
Jak znaleźć API Endpoint

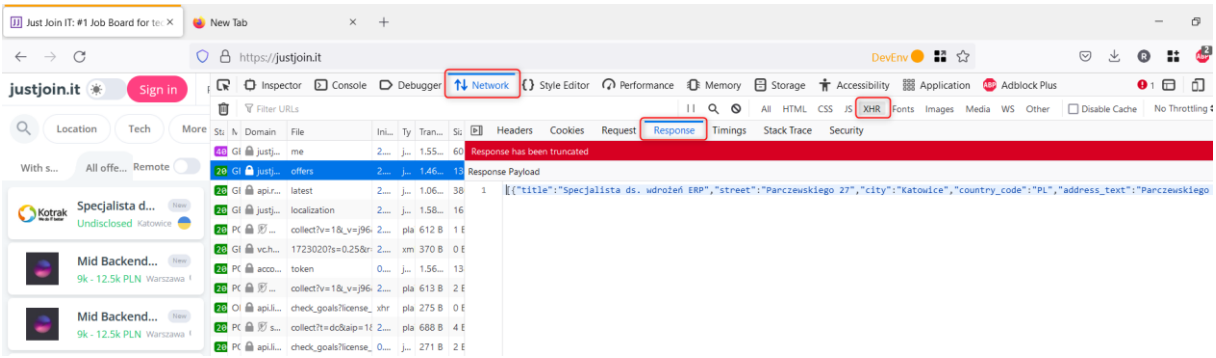
W przeglądarce (screeny pochodzą z Firefox ale w Chrome wygląda to podobnie) odpal *Developer Tools* (np. F12 lub z menu przeglądarki, zazwyczaj w opcji *More Tools*) i wpisz adres interesującej Cię strony. Chcesz zobaczyć requesty jakie są wysyłane i pobierane przez daną stronę. Interesuje nas zakładka *Network* a w niej głównie *XHR* (XMLHttpRequest) – opcje zaznaczone poniżej.



Jak wyłuskać ten właściwy request?

W gąszczu requestów, które pojawią się w zakładce *Network* interesują nas oczywiście tylko te, które pochodzą z oryginalnej strony (zwróć uwagę na kolumnę *Domain*).

W *Headers* zakładce *XHR* sprawdzamy, czy mamy jakiś sensowny URL i czy dany request zwraca jakieś dane ustrukturyzowane:

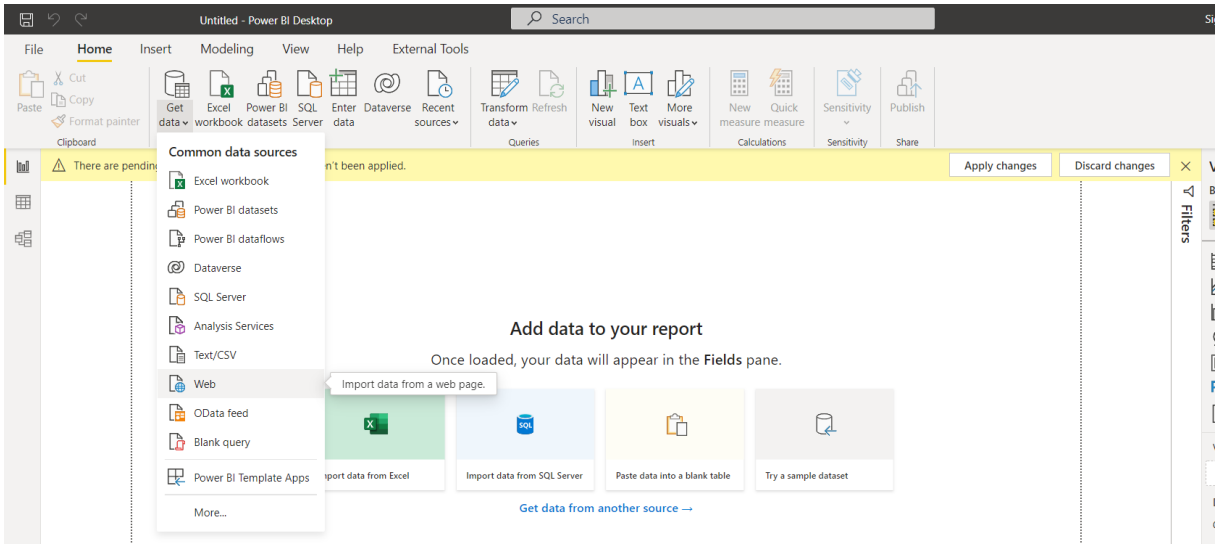


Jeśli tak i te dane wyglądają na to, czego poszukujemy, to kopiujemy URL z zakładki *Headers* i możemy od razu przejść do Power BI.

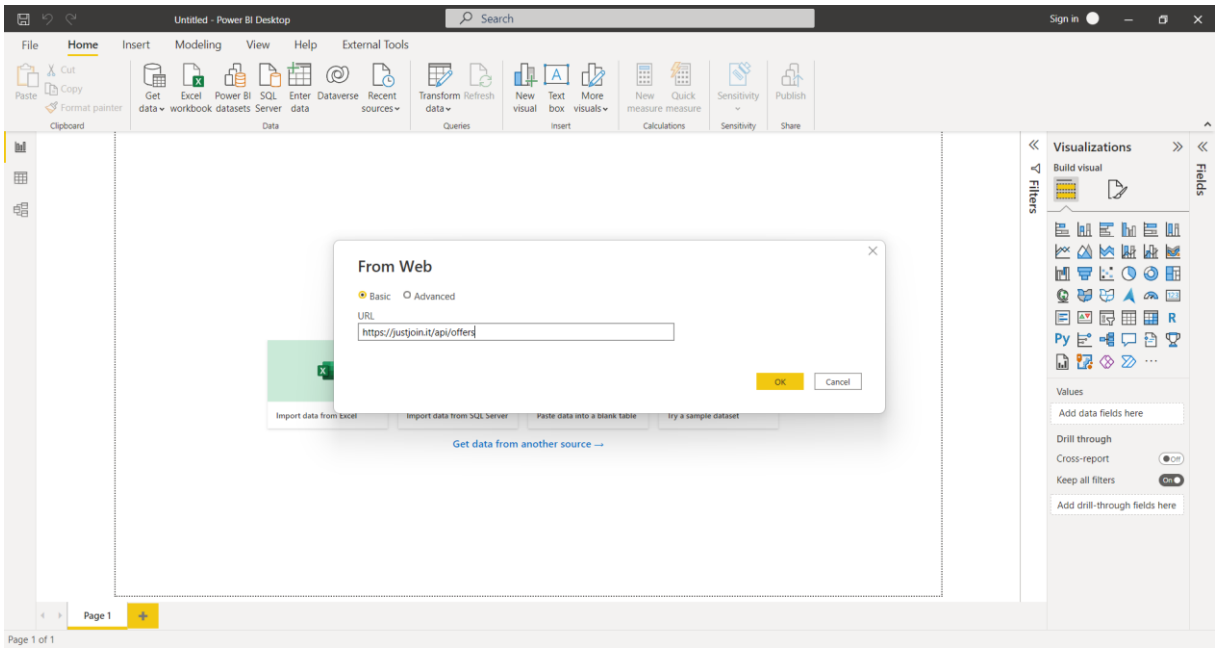
Zazwyczaj trzeba trochę pogrzebać i poprzyglądać się, który request będzie tym, którego potrzebujemy. Jeśli nie widzimy niczego ciekawego, albo nie ma requestów, to czasem warto odświeżyć lub wejść w pojedynczy rekord i na nim się przyglądać jak wygląda URL (możemy potrzebować tylko części albo tez dany URL zwraca nam wynik przefiltrowany, a my możemy odpytać o całość).

Ściąganie danych do Power BI

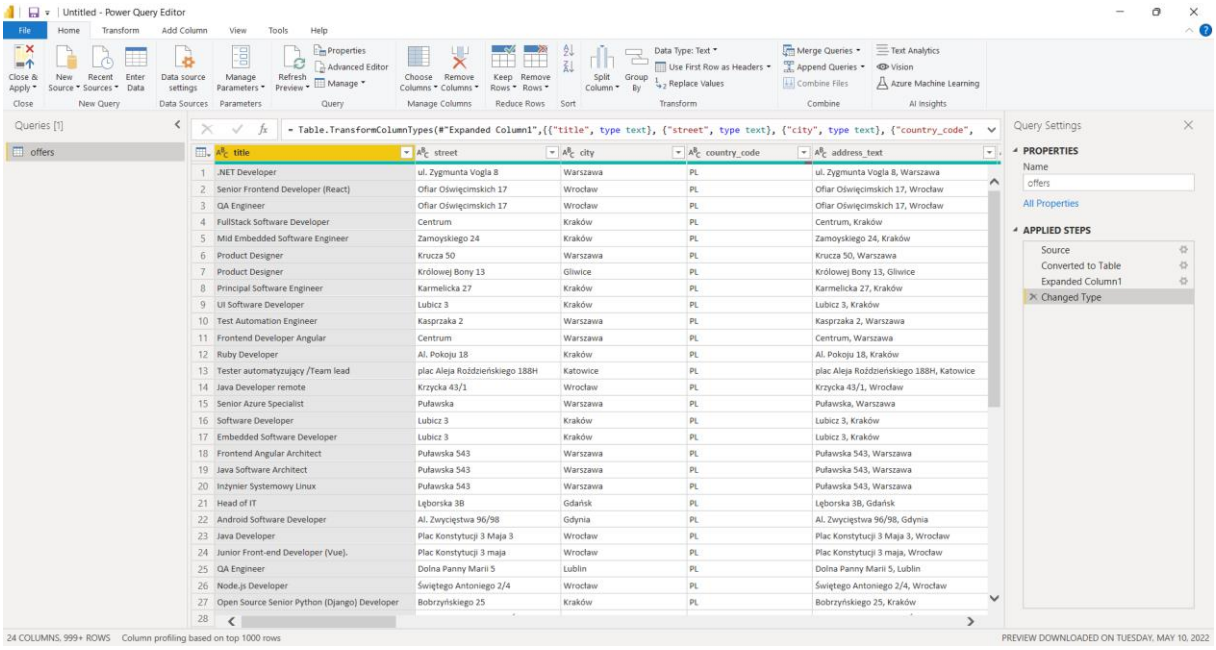
W Power BI korzystamy z konektora Web:



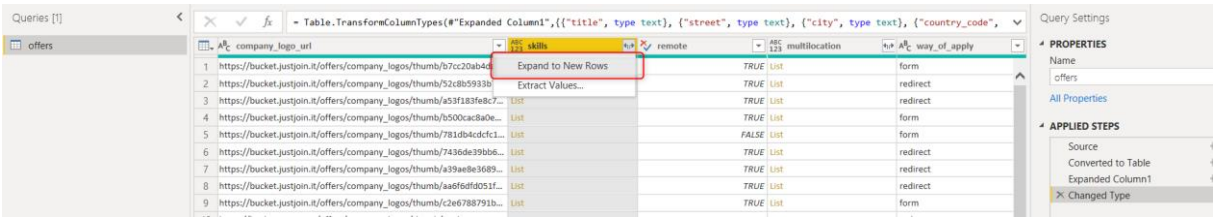
I po prostu wklejamy nasz URL:



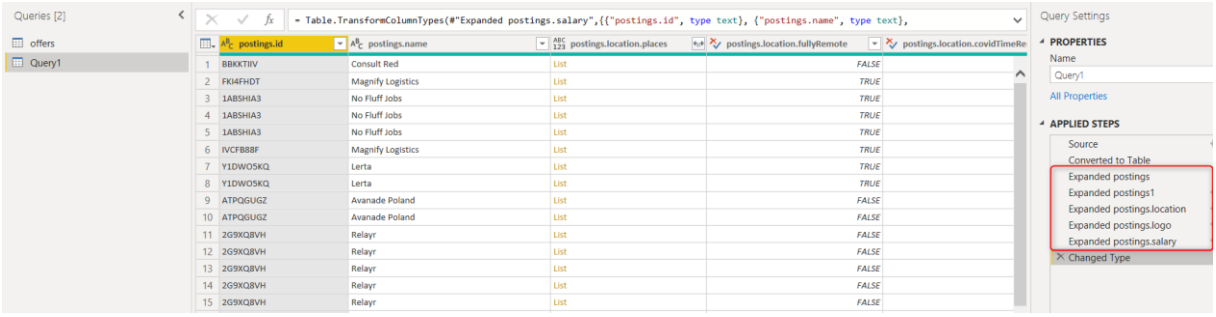
Sam Power BI jest na tyle sprytny, że poradzi sobie z przetwarzaniem danych ustrukturyzowanych z pobranego URLa, więc naszym zadaniem jest głównie sprawdzenie, czy transformacje są takie, jakich chcieliśmy i czy potrzebujemy czegoś więcej (lub mniej).



Polecam przejrzenie po kolei tego co się dzieje w poszczególnych krokach i zwrócenie uwagi na to, że bardzo często pobieramy listy rekordów, a niejednokrotnie interesujące nas kolumny to znowu rekordy lub listy, więc je rozwijając będziemy duplikować wiersze:



Pamiętaj też, że Power BI lubi czasem zrobić za nas trochę za dużo 😜 więc uważaj na kroki, które dorzuca automatycznie. Zawsze możesz je powyrzucać (i warto się tych nadmiarowych pozbywać), np. tu:



No to teraz przed Tobą mnóstwo świetnej zabawy 😊 Przeróbki w Power Query, łączenie danych, dorzucanie funkcji, tworzenie prawidłowego modelu itp. Itd. (czyli to co tygryski lubią najbardziej).

Przy tworzeniu raportu ograniczają nas już tylko nasza wyobraźnia (no i trochę też ograniczenia techniczne).

Pamiętajmy o tym, że tak pozyskane dane pokazują nam stan „na dzisiaj”, więc gdybyśmy chcieli porównywać je w czasie, to należałoby je też w jakiś sposób odkładać. Do pełni szczęścia brakuje nam więc jakiegoś ekstraktu tych danych (najlepiej w sposób oczywiście automatyczny).

Ekstrakt danych z Power BI do pliku płaskiego

Kiedy skończysz już przerabianie swoich danych i masz gotowe query wynikowe, możesz pokusić się o eksport tych danych. Do zastanowienia się jest oczywiście kwestia, które dane będą wymagały takiego ekstraktu, czy np. tylko tabela faktów, czy także część słowników. W moim przypadku najprostszym sposobem wydał mi się skrypt R.

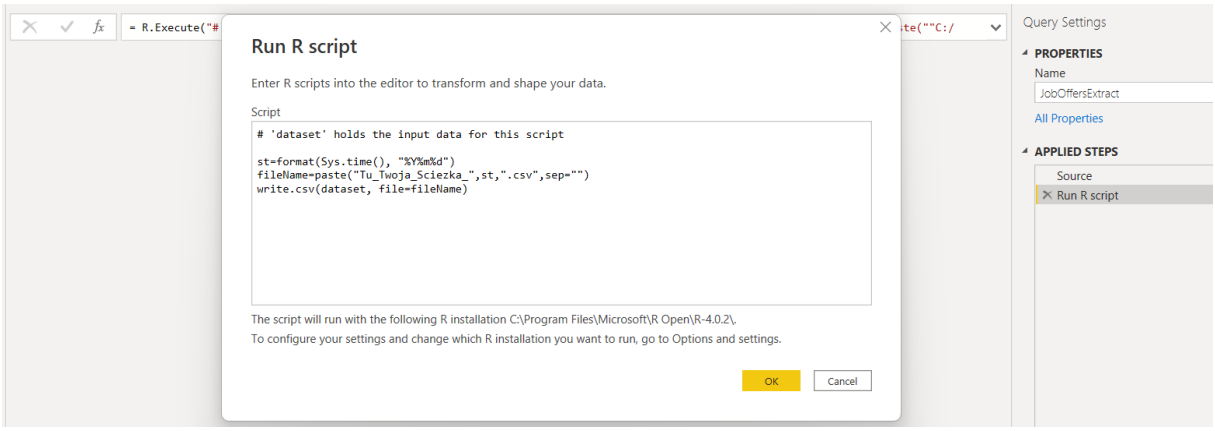
Z wymagań, które muszą być spełnione przed tym krokiem, to musisz mieć zainstalowany R (tego kroku nie opisuję, odsyłam do dokumentacji).

Na gotowym query wynikowym (tym lub tymi, które chcesz eksportować), a dokładniej kopii tego query (używam zarówno widoku bieżącego jak i tych danych do eksportu, więc potrzebuję dwóch query) dorzucamy krok *Run R Script*:

Najprostszy ekstrakt, który zadziała to:

```
write.csv(dataset, file = "Twoja_Sciezka.csv")
```

Ale ponieważ nam zależy na schedulowanym ekstrakcie (czyli danych zrzucanych np. każdego dnia), to zrobimy odrobinę bardziej skomplikowany skrypt:



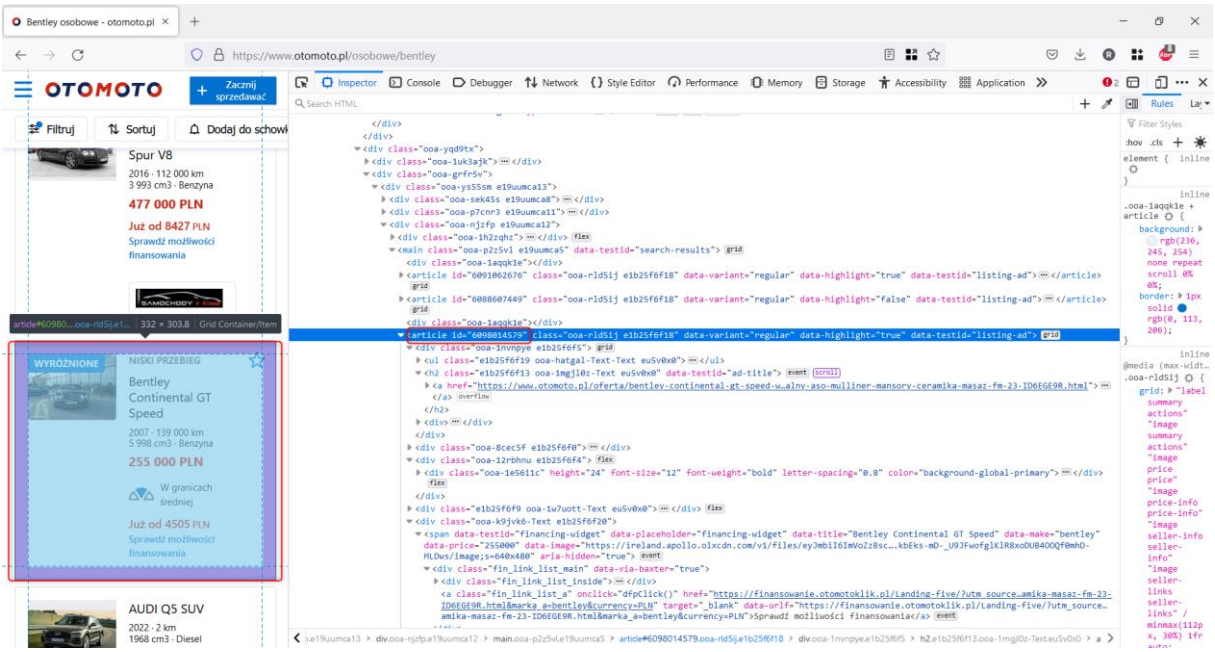

```
st=format(Sys.time(), "%Y%m%d")

fileName=paste("Twoja_Sciezka ",st,".csv",sep="")

write.csv(dataset, file=fileName)
```

I tyle już wystarczy 😊

Kolejnym krokiem, aby działo się to automatycznie jest ustawienie harmonogramu. Tu posługujemy się Data Gatewayem w personal mode (niestety na ten moment działa to z personal mode). Instalacja i harmonogram są już standardowe, więc nie będę się nad tym rozwodzić.

Wyłuskiwanie danych z tagów

Nie zawsze jednak uda nam się znaleźć adres endpointa. W tym przypadku pozostają nam stare dobre tagi i znajdowanie wzorców pośród nich, a następnie dzielenie poszczególnych sekcji w Power Query. W tym przypadku najprościej skorzystać znowu z Developer Tools (F12) i tym razem z zakładki Inspector. W poniższym przykładzie możemy zauważyć, że np. każda nowa oferta rozpoczyna się od tagu *article*.



Dalsze kroki będą już analogiczne, czyli dużo Power Query, dużo transformacji aż do momentu, gdy jesteśmy zadowoleni z uzyskanego efektu.

O czym pamiętać

Na koniec jeszcze jedna informacja, na którą należy zwracać uwagę, czyli czy w ogóle możemy scrapować dane. Pamiętamy o tym, żeby sprawdzić strony, z których chcemy pobierać dane, nie naruszać praw autorskich i ogólnie przestrzegać regulaminów. Polecam pogooglować bo to kolejny bardzo interesujący temat.