Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Hurtownie danych i systemy Business Intelligence

Eurowizjanie - projekt

Autorzy:
Jakub Kosterna, Patryk Wrona

Prowadzący laboratorium:
mgr inż. Jakub Abelski

Warszawa
16 maja 2021

Spis treści

1. Cel projektu oraz korzyści z perspektywy odbiorcy

1.1. Cel projektu

Konkurs Piosenki Eurowizji to nie tylko największe muzyczne wydarzenie roku, które od dekad śledzą setki milionów telewidzów, ale także i bardzo interesujący event pod kątem analizy biznesowej. Na przełomie lat systemy głosowań i wyboru zwycięzcy ewoluowały, w efekcie obfitując w naprawdę pokaźne dane. Nie trudno znaleźć w Internecie chociażby tabelę prawie 50 000 obserwacji, zawierającą kompletne informacje o wynikach ostatnich 46 lat.

Nietrudno zauważyć ciekawe własności konkursu i prawa, jakim się on rządzi. Niezaprzeczalnym faktem jest zjawisko częstszego głosowania między sympatyzującymi między sobą kulturowo państwami, a i forma w poszczególnych latach ma kolosalny wpływ na rezultaty. Konkurs jest także niezwykle godny uwagi pod względem socjologicznym - eksperci rok w rok przewidują, która 3-minutowa propozycja tym razem zyska najwięcej aprobaty u telewidzów, a która - u profesjonalnego jury.

Cel projektu jest zbadanie danych dotyczących Konkursu Piosenki Eurowizji i wyciągnięcie z niego wartościowych i nieoczywistych wniosków.

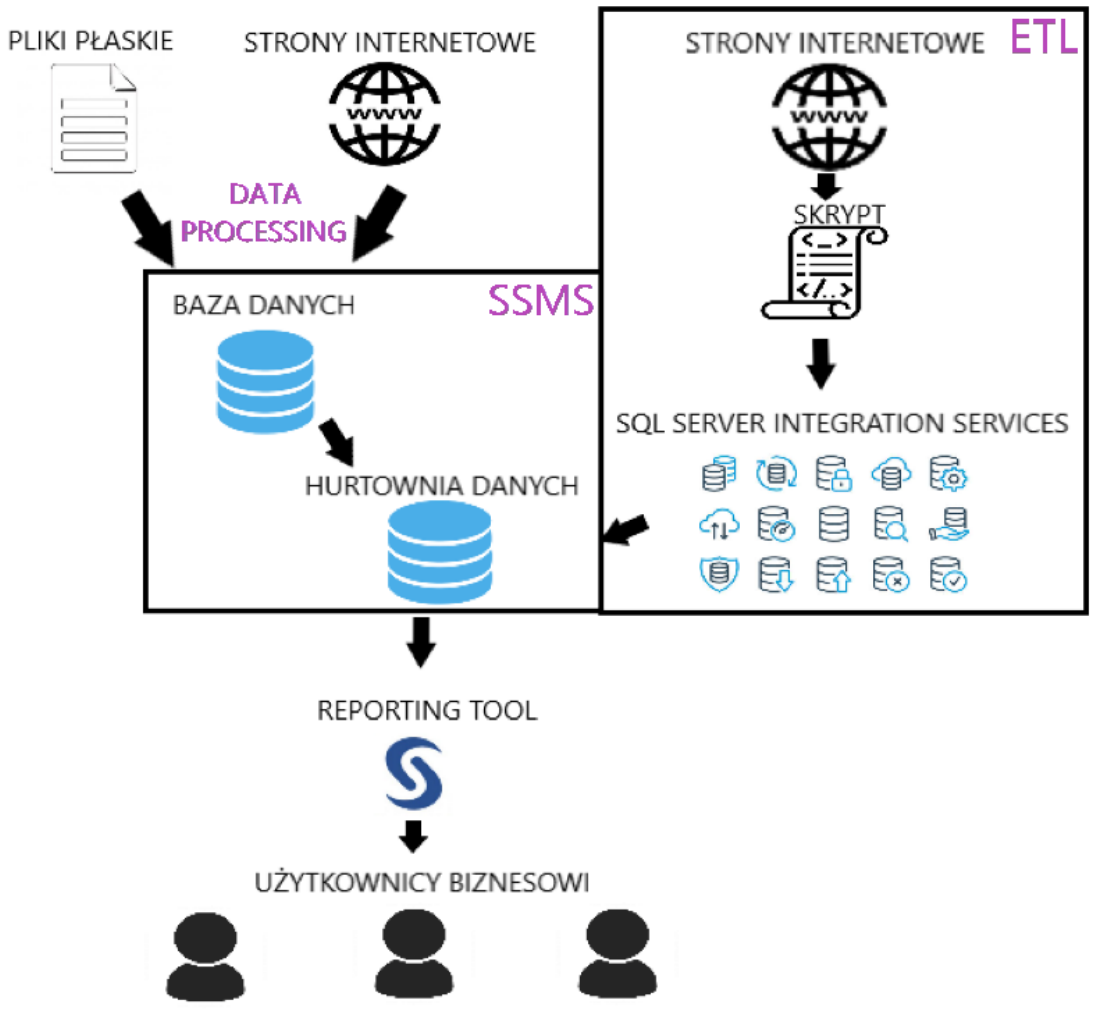
1.2. Korzyści z perspektywy odbiorcy rozwiązania

Poznając Eurowizję tak naprawdę poznajemy jak działają masy w rozrywkowe wieczory przed telewizorem. Konkurs Piosenki Eurowizji jest jedynym takim wydarzeniem muzycznym, gdzie biorą udział reprezentanci tak wielu różnych kultur, a rok w rok o głównym wyniku w dużej mierze decydują ludzie w tak dużej liczbie - upragnionych do głosowania jest często blisko miliard osób!

Bez wątpienia korzyścią z zaproponowanej architektury jest wartościowe poznanie człowieka XX i XXI wieku - przynajmniej pod kątem spontanicznych sympatii do artystycznych mini-dzieł jakimi są utwory KPE (*Konkursu Piosenki Eurowizji*). Opracowane współczynniki i własności, jakie oferuje rozwiązanie, są także bardzo pomocne pod względem predykcji przyszłych wyników - co przełożyć się może finansowo chociażby przy okazji taktycznego podejścia do zakładów bukmacherskich.

2. Diagram i opis planowanej architektury rozwiązania

2.1. Diagram architektury



Rys. 2.1. Diagram architektury

2.2. Opis architektury

Dane ze stron internetowych oraz z plików płaskich stworzonych na podstawie znalezionych dodatkowych informacji przetworzono i umieszczono w bazie danych za pomocą SQL Server Management Studio. W ramach procesu ETL wykonuje się skrypt(Web Scraping script) pobierający dane ze strony internetowej i dzięki SQL Server Integration Srvices dane są odpowiednio przetwarzane oraz po zapewnieniu, że mają odpowiednie dla bazy danych typy i nie naruszają klucza głównego, są one umieszczane w bazie danych. Baza danych jest modelowana na zasadzie modelu gwiazdy - staje się hurtownią danych. Następnie dane z hurtowni danych są przygotowywane do analizy poprzez system raportowania SAS, który finalnie mogą używać użytkownicy biznesowi.

3. Opis wykorzystanych zbiorów danych

W naszym projekcie wykorzystaliśmy poniższe zbiory danych:

3.1. Informacje o uczestnikach Eurowizji [1]

Tabela zawiera dane o utworach zgłoszonych na konkurs. Można w niej znaleźć bardziej oczywiste fakty, takie jak tytuły utworów i pseudonimy / nazwiska ich wykonawców; ale także takie kolumny jak tekst utworu (u nas uproszczony do długości w znakach), nazwiska kompozytorów, pozycje startowe na kolejnych etapach konkursu czy wyniki w końcowych rankingach.

3.2. Informacja o przyznanych punktach przez poszczególne państwa [2]

Jest to główna i jedyna tabela faktów. Zawiera ona najwięcej - bo prawie 50 000 - wierszy zawierających fakty historyczne o przyznawanych punktach między państwami. Rozróżnione są punkty od jurorów i telewidzów (w latach, gdzie były one liczone osobno) oraz etapy konkursu - ewentualne półfinały oraz finał. Ramka jest także źródłem dla tabeli *coefficients*.

3.3. Informacje o jury Eurowizji [3]

Dane pobierane bezpośrednio z internetowego API - przyjaznego projektom takim jak ten **Data World** - https://data.world/rhubarbarosa/eurovisionvotingstats (ostatni plik .csv). Na cele hurtowni został on okrojony i nieco przerobiony - dla kazdego państwa i roku mamy informacje o liczbie mężczyzn oraz kobiet w jury, a także ich średniej wieku.

3.4. Dane dotyczące PKB krajów [4]

Dane dotyczące wartości PKB oraz procentowego wzrostu PKB wielu krajów (w tym również krajów biorących udział w Eurowizji) na przestrzeni lat 1960-2020. Przetworzone do pliku płaskiego ładowanego do bazy danych SQL Server - przy okazji uwzględniając tylko lata odpowiadające latom, w których organizowany był Konkurs Piosenki Eurowizji.

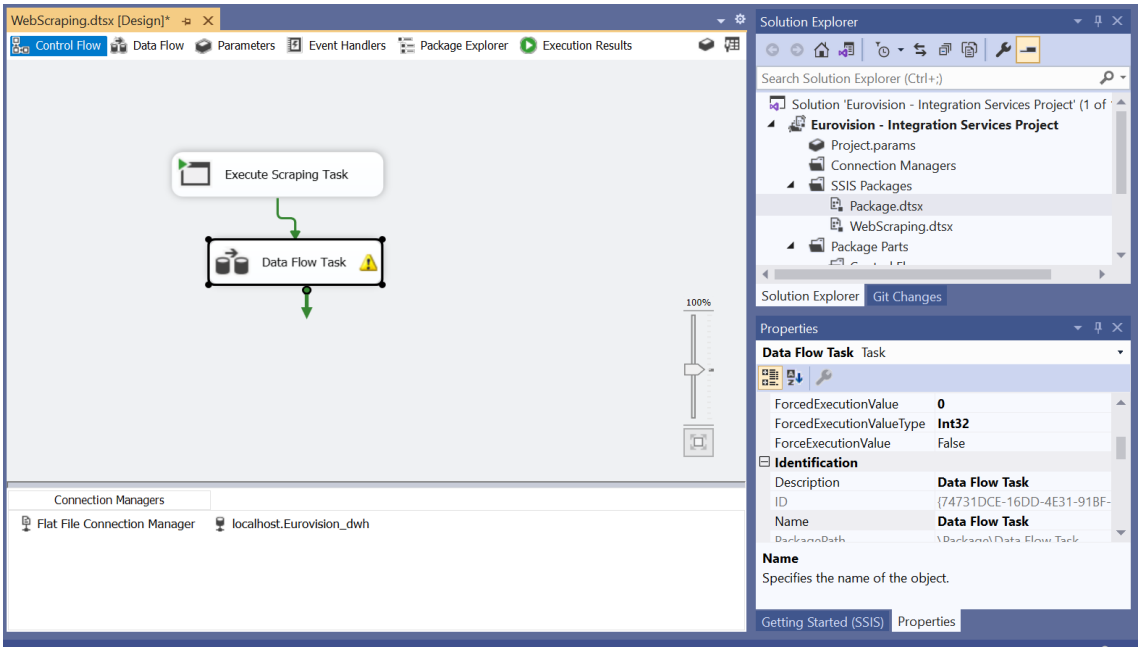
3.5. Wymiar geograficzny

To jedyna taka tabela utworzona w pełni ręcznie - dla 52 krajów biorących udział w KPE na przełomie analizowanych lat zostały dobrane 3 wymiary - część kontynentu (np. Azja Wschodnia), kontynentu (zwykle jeden kontynent, ale niekiedy chociażby Europa Azja - w przypadku krajów leżących na dwóch kontynentach) i wreszcie kontynent. W przypadku rejonów dyskusyjnych tereny zostały przypasowane po uzgodnieniu przez pełen zespół. Obowiązują podstawowe zasady takie jak fakt spójności grup części kontynentów (połączone tworzą jeden ląd), ale w niektórych przypadkach sprawa była nieoczywista i jednoznaczne wybory mogą niekiedy być nieco kontrowersyjne.

Uwaga! Ze względu na duże braki danych i w znalezionych tabelach, i biorąc pod uwagę oficjalne komunikaty organizatorów - analizowane są tylko KPE od roku 1975. w górę.

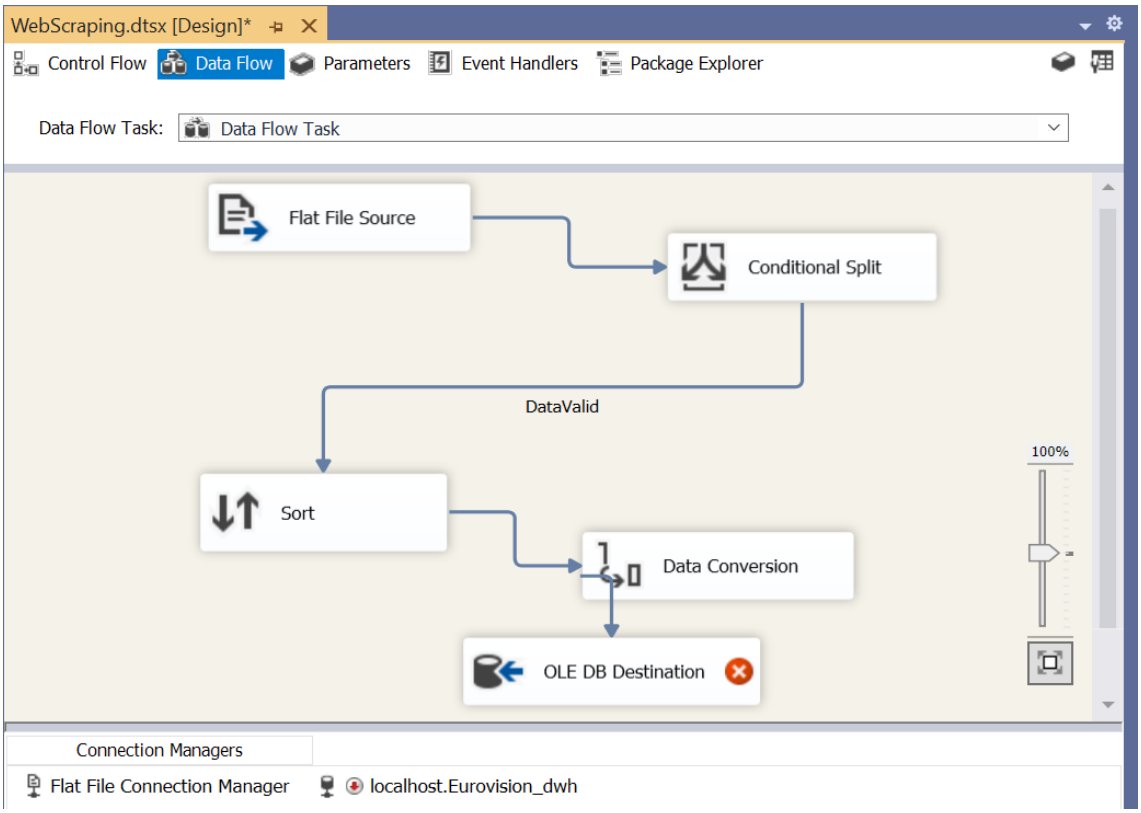
4. Transformacje danych w procesach ETL

Proces ETL wykorzystuje w swoim Control Flow następujące zadania:



Rys. 4.1. ETL - Control Flow

Korzystając ze źródła dotyczącego Jury [3], proces ETL wykorzystuje w module **Execute Scraping Task** skrypt napisany w języku R, który pobiera dane ze strony internetowej, przetwarza je i tworzy plik. Nastepnie plik pobierany jest jako **Flat File Source**, a schemat przetwarzania danych w procesie ETL wygląda następująco:



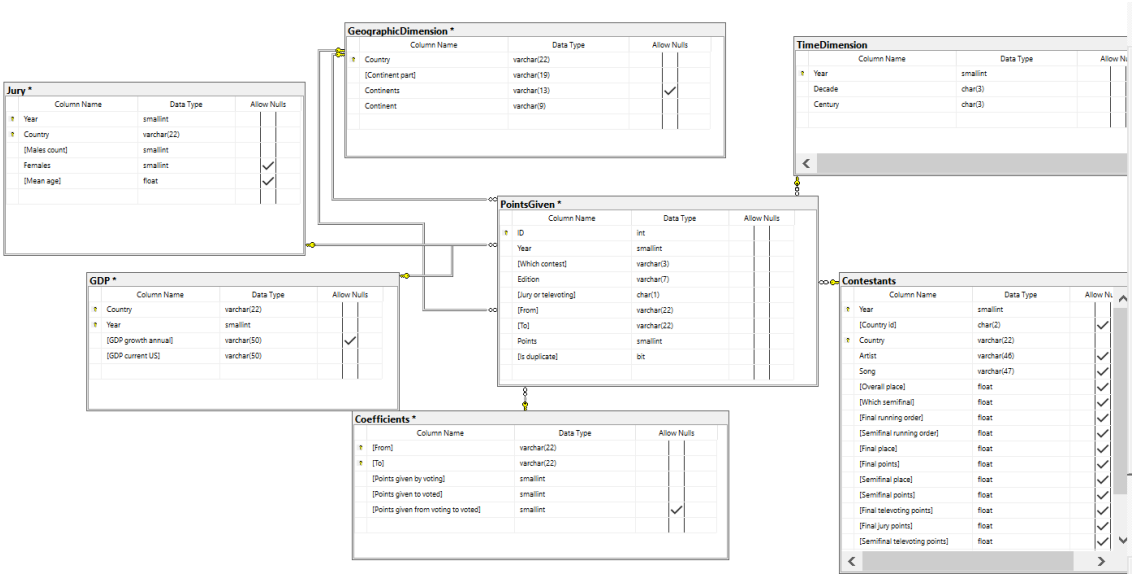
Rys. 4.2. ETL - Data Flow

Poszczególne komponenty realizują poniższe zadania:

— **Flat File Source** - pobiera dane z pliku (w tym przypadku o rozszerzeniu .csv), tworzonego przez skrypt pobierający te dane ze strony internetowej. Wymaga Connection Managera.
— **Conditional Split** - jeśli w modelu hurtowni danych dane kolumny nie moga być puste, nie przekaże dalej takich obserwacji
— **Sort** - zapewnia odpowiednie sortowanie kolumn w hurtowni danych oraz uniemożliwia wystąpienie ewentualnych duplikatów
— **Data Conversion** - zamienia typy danych poszczególnych kolumn, zazwyczaj wymagały pszekształcenia z non-unicode na unicode w celu umieszczenia ich w hurtowni danych
— **OLE DB Destination** - umieszcza dane w lokalnej hurtowni danych wiersz po wierszu.Wymaga Connection Managera.

5. Model hurtowni danych wraz z opisem komponentów

Stworzony przez nas model hurtowni danych jest modelem gwiazdy:



Rys. 5.1. Model hurtowni danych - gwiazda

W modelu zakładamy istnienie danych komponentów:

— **PointsGiven** - Tabela faktów. Zawiera informacje o nadaniu liczby punktów (kolumna Points) przez dane państwo (kolumna From) innemu państwu (kolumna To). Zawiera znacznik czasowy będący rokiem - kolumnę Year oraz informację o tym czy był to głos nadany przez jury czy poprzez teleglosowanie (kolumna [Jury or televoting]). Kolumna "which contest" zawiera informację czy był to finał lub któryś w półfinałów. Kolumna "Edition" jest połączeniem Year oraz kolumny "which contest".
— **Jury** - Tabela wymiarów. Zawiera informacje dotyczące każdego państwa w danym roku Eurowizji(jest to złożony klucz główny Year oraz Country). Dane te obejmują liczbę meżczyzn i kobiet wchodzących w skład jury tego państwa oraz ich średni wiek.
— **GeographicDimension** - Tabela wymiarów. Hierarchia geograficzna. Każdemu państwu biorącemu udział w Konkursie Piosenki Eurowizji (klucz główny) odpowiadają kolumny Continents(jeśli terytorium państwa jest na kilku kontynentach), wybrany główny kontynent - Continent, oraz cześc kontynentu (Continent Part) np. Europa Zachodnia.
— **TimeDimension** - Tabela wymiarów. Wymiar Czasowy. Zawiera informacje dla danego roku, w której jest dekadzie i stuleciu.
— **Contestants** - Tabela wymiarów. Kluczem głównym jest oczywiście rok i kraj - dla każdej pary mamy informację o uczestniku - imię, tytuł piosenki, miejsce z którego startował i które uzyskał oraz poszczególne punkty, które zawdzięcza jury oraz teleglosowaniu.

— **GDP** - Tabela wymiarów. Zawiera ona kolumny dotyczące wartości oraz procentowego wzrostu PKB dla danego państwa w danym roku. Kluczem głównym jest oczywiście państwo(Country) oraz rok(Year).

— **Coefficients** - to jedyna taka tabela wymiarów - opisuje współczynniki sympatii między państwami. Zawiera +2000 par państwami, które mają taką właściwość, że państwo B przynajmniej raz wzięło udział w finale, w którym A było uprzywilejowane do głosowania (czyli po prostu brało udział w danym roku). Dla każdej pary został obliczony wspomniany wcześniej współczynnik sympatii(kolumna [Points given from voting to voted]) - jest to różnica procenta punktów, które państwo A przekazało państwu B w opisanych wcześniej sytuacjach (oba kraje w grze, B w finale) i średniego procenta głosów uzyskiwanych w finale przez państwo w owych latach. Owa miara naturalnie przekłada się na realny stosunek między uczestnikami na Eurowizji, niezależnie od ich typowych wyników (gorszych czy lepszych) oraz lat udziału. Wynikowo wynik bliski zera oznacza neutralny stosunek, dodatni - większy niż typowy, zaś ujemny - nieprzychylność. Tak oto można zaobserwować ogromną sympatię między Grecją a Cyprem, zaś z drugiej strony absolutną niechęć do siebie między Armenią i Azerbejdżanem. Co ciekawe, Polska najlepszy współczynnik ma od Niemców. Może to lekko zaskoczyć, ale jest bardzo sensowne - Polacy obejmują tam dużą część imigrantów i społeczność Niemiec ogólnie ma dosyć nikłe zainteresowanie konkursem - stąd średnio bardziej zaangażowani polacy głosują na swoich.
  Tabela współczynników sympatii została skonstruowana w języku C++ z danych poczerpanych z głównej tabeli faktów i załadowana do hurtowni danych.

6. Opis planowanej warstwy raportowej

Warstwa raportowa będzie wykorzystywała system SAS Visual Analytics. Docelowo posłuży ona do wykonania ciekawych wizualizacji i raportów zawierających godne uwagi dane uzyskane z pomocą hurtowni i procesu ETL.

W zależności od możliwości narzędzia i pomyślności domniemanych rozwiązań, zostaną przedstawione takie informacje jak:

- wizualizacje dotyczące typowych różnic w ostatecznym rankingu 10 najlepiej ocenionych utworów
- wykresy okazujące różnicę "sympatii" państwa do innych - na przykład w postaci słupków przedstawiających częstość głosowania na państwo odnoszącą się do długości odpowiadającej mu kolumnie
- zestawienia płci i wieku jurorów i obserwacja trendów na przełomie lat
- fakty dotyczące miar centralnych pozycji końcowych wybranych krajów w ustalonym okresie czasowym
- porównanie własności cech składu jury dla konkretnych państw i sympatii do poszczególnych utworów (na przykład przybliżenie odpowiedzi na pytanie, czy kobiety częściej faworyzować będą inne kobiety albo sympatię głosujących częściej otrzymują utwory reprezentantów w podobnym wieku)

Cały proces w wielkim skrócie zostanie zrealizowany w czterech krokach:

1. Wybrane tabele i segmenty ETL zostaną wprowadzone z pomocą modułu *Data Preparation*
2. Dodatkowe ciekawe zależności zostaną znalezione i następnie uwzględnione przy przeszukiwaniu *Data Explorer*
3. Na podstawie zebranych danych i dostępnych możliwości zostanie obmyślony i przygotowany raport wynikowy poprzez *Raport Designer*
4. Wynik pracy zostanie zweryfikowany, a następnie ewentualnie ulepszony po analizie *Raport viewer*.

Bibliografia

[1] Informacje o uczestnikach Eurowizji - contestants.csv https://github.com/Spijkervet/eurovision-dataset/releases

[2] Informacja o przyznanych punktach przez poszczególne państwa https://data.world/datagraver/eurovision-song-contest-scores-1975-2019

[3] [Dane pobierane z internetu przez proces ETL] Informacje o jury Eurowizji - VotingJury.csv https://data.world/rhubarbarosa/eurovisionvotingstats

[4] Dane dotyczące PKB krajów https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[5] Lista państw biorących udział w Eurowizji - dane przetworzone ze strony https://en.wikipedia.org/wiki/List_of_countries_in_the_Eurovision_Song_Contest