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
8 czerwca 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.

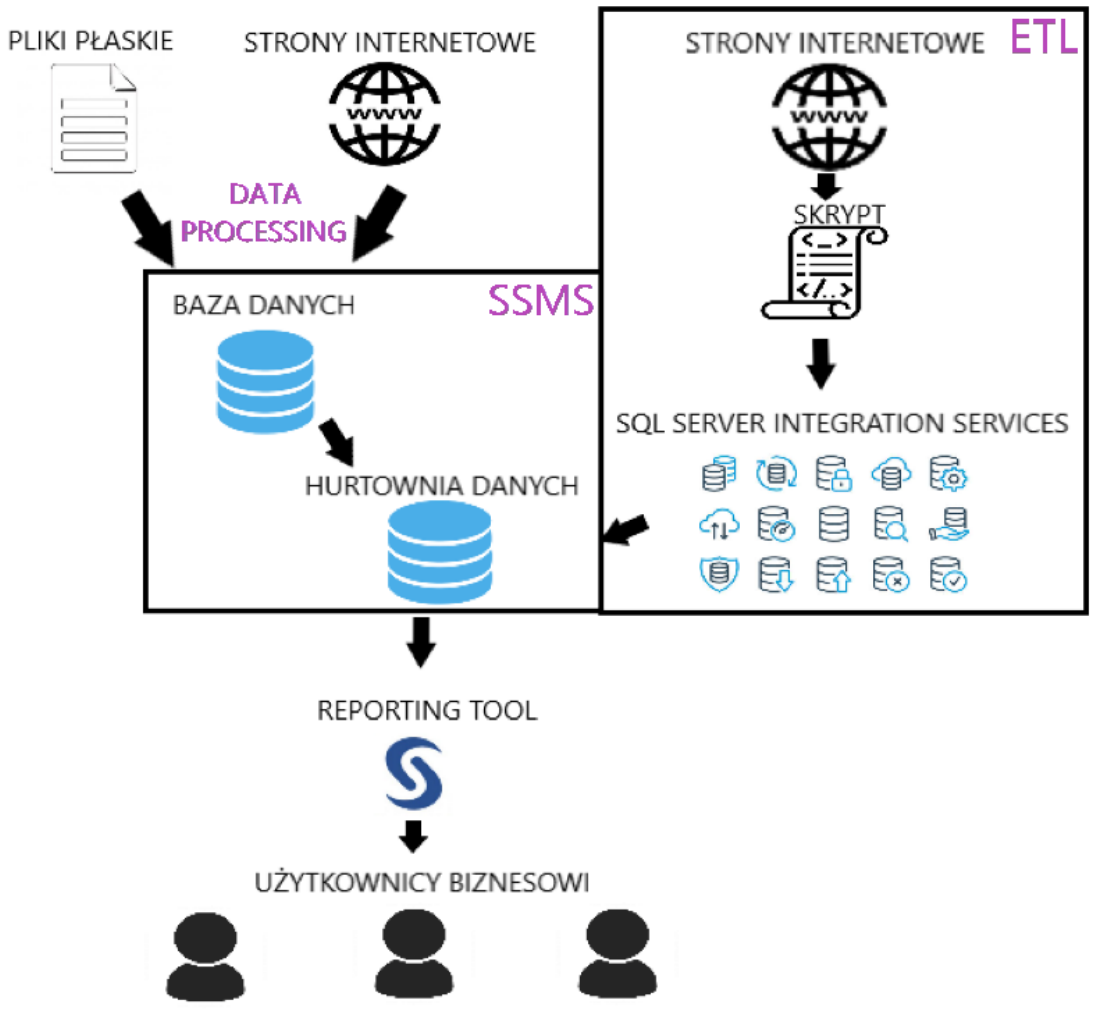
1.2. Korzyści z perspektywy odbiorcy rozwiązania

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]

Sa to dane postawowe do tabeli faktów, aktualizowane co roku. Pobierane one są z internetowego API w trakcie procesu ETL. Zawiera ona najwięcej - bo ponad 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.

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 każdego 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

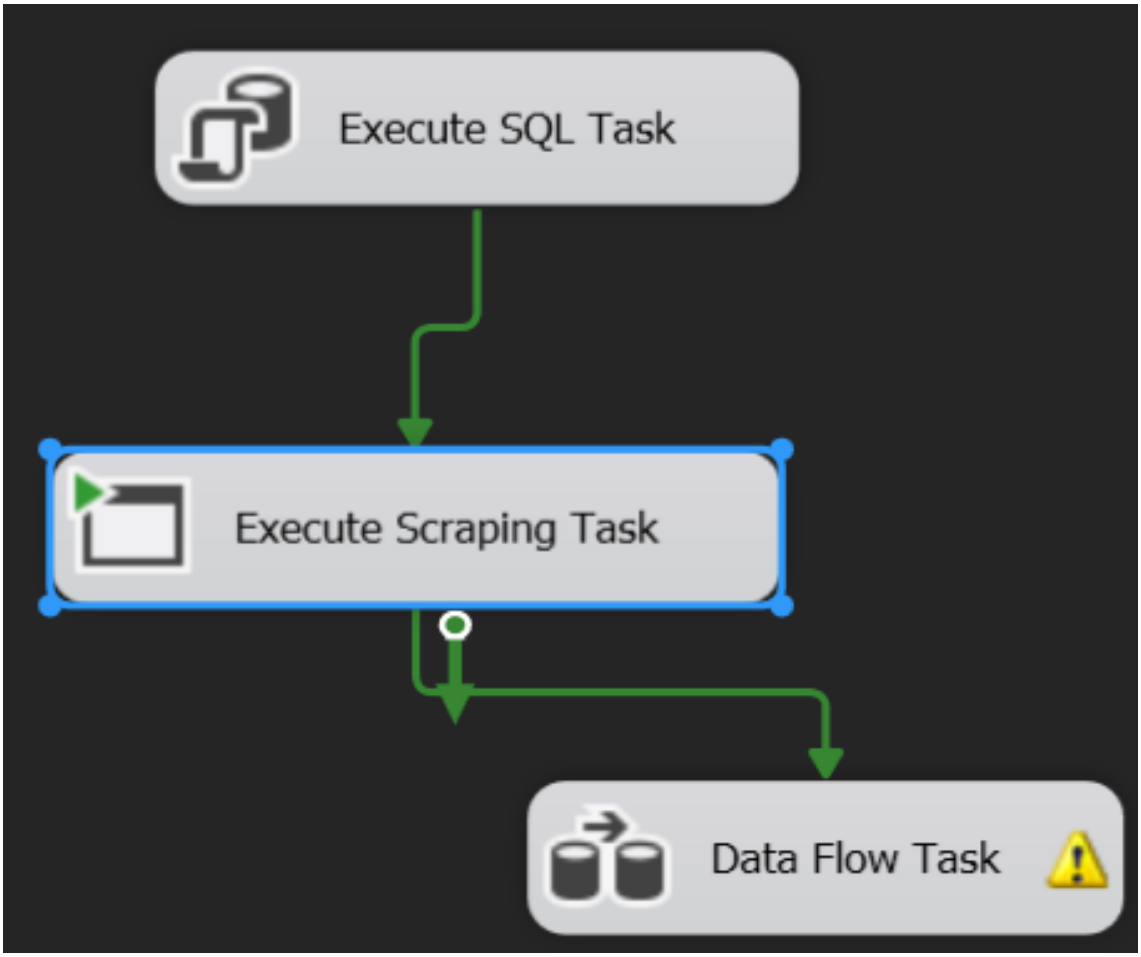
To jedyna taka tabela utworzona w pełni ręcznie - dla 52 krajów biorących udział w KPE na przełomie analizowanych lat zostały dobrane 3 wymiary - część kontynentu (np. Azja Wschodnia), kontynentu (zwykle jeden kontynent, ale niekiedy chociażby Europa Azja - w przypadku krajów leżących na dwóch kontynentach) 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

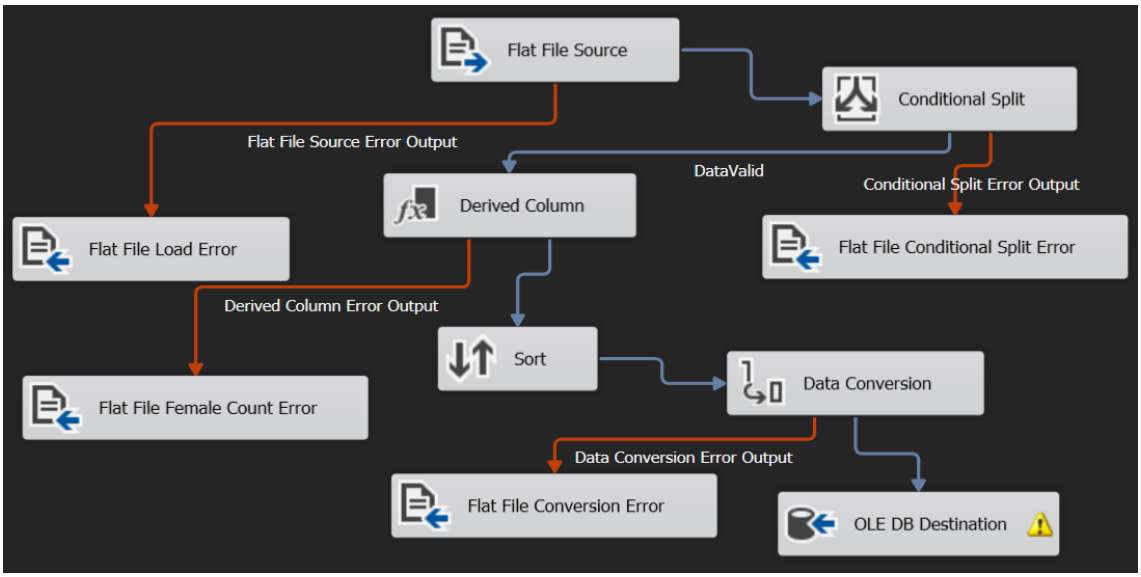
Procesy ETL wykorzystują w swoim Control Flow następujące zadania:

4.1. ETL - Jury



Rys. 4.1. ETL - Jury Control Flow

Na pierwszym miejscu uruchamiany jest skrypt w ramach zadania **Execute SQL Task**, który tworzy na nowo odpowiednią tabelę Jury, w której będzie można dodać aktualne dane dotyczące Jury. Po drugie, korzystając ze źródła dotyczącego Jury [3], proces ETL wykorzystuje w module **Execute Scraping Task** skrypt napisany w języku R, a uruchamiany poprzez skrypt .bat, który pobiera dane ze strony internetowej, przetwarza je i lokalnie tworzy plik .csv. Następnie plik pobierany jest jako **Flat File Source**, a schemat przetwarzania danych w procesie ETL wygląda następująco:

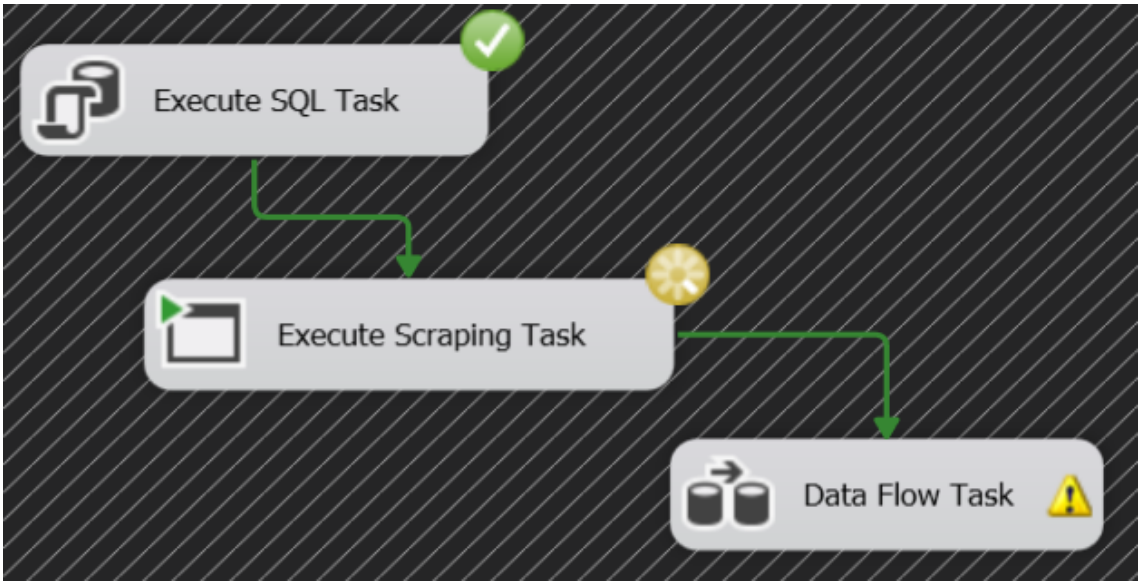


Rys. 4.2. ETL - Jury Data Flow

Poszczególne komponenty realizują poniższe zadania:

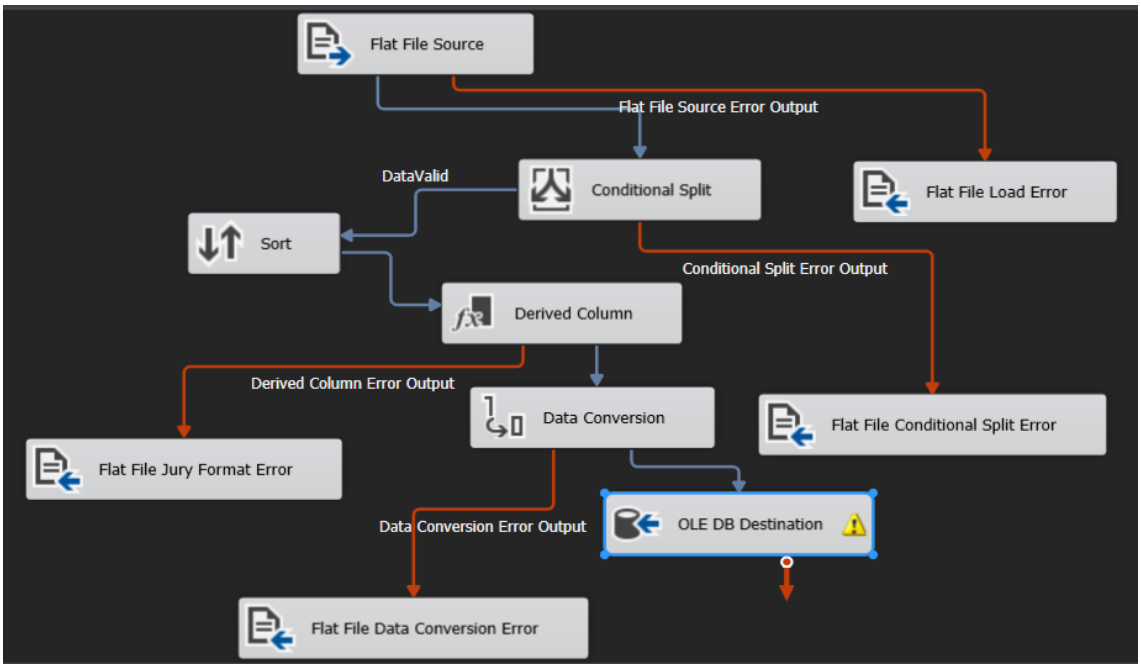
— **Flat File Source** - pobiera dane z pliku (w tym przypadku o rozszerzeniu .csv), tworzonego przez skrypt pobierający i przetwarzajacy te dane ze strony internetowej. Wymaga Connection Managera.
— **Conditional Split** - jeśli w modelu hurtowni danych dane kolumny nie moga być puste, nie przekaże dalej takich obserwacji.
— **Derived Column** - tworzy nowe kolumny pochodzące od już istniejących kolumn - nową kolumnę 'Female count' oraz zmienia już istniejącą 'Jury or televoting'. Female count jest obliczany z Male count oraz informacji, że w skład Jury zawsze wchodzi 5 osób.
— **Sort** - zapewnia odpowiednie sortowanie kolumn(po kraju Jury, a następnie po roku konkursu) 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.

4.2. ETL - Points Given



Rys. 4.3. ETL - Points Given Control Flow

Na pierwszym miejscu uruchamiany jest skrypt w ramach zadania **Execute SQL Task**, który tworzy na nowo odpowiednią tabelę Points Given, w której będzie można następnie dodać aktualne dane dotyczące punktów nadanych przez poszczególne państwa w konkursie Eurowizji. Po drugie, korzystając ze źródła dotyczącego wyników Eurowizji [2], proces ETL wykorzystuje w module **Execute Scraping Task** skrypt napisany w języku R, a uruchamiany poprzez skrypt .bat, który pobiera dane ze strony internetowej, przetwarza je i tworzy lokalnie plik .csv. Następnie plik pobierany jest jako **Flat File Source**, a schemat przetwarzania danych w procesie ETL wygląda następująco:



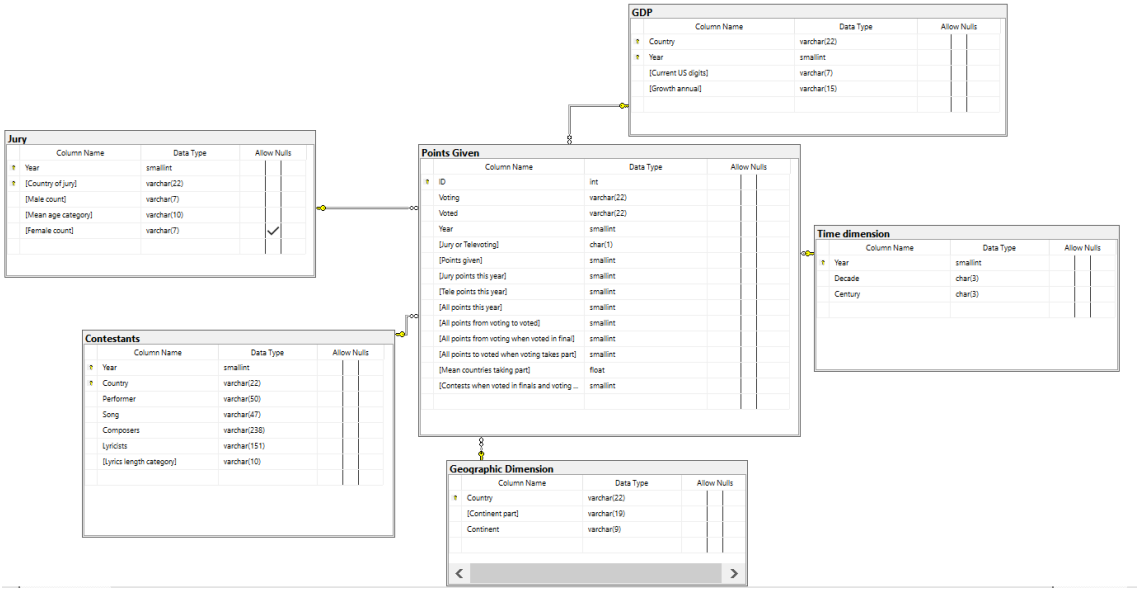
Rys. 4.4. ETL - Points Given Data Flow

Poszczególne komponenty realizują poniższe zadania:

— **Flat File Source** - pobiera dane z pliku (w tym przypadku o rozszerzeniu .csv), tworzonego przez skrypt pobierający i przetwarzajacy 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.
— **Derived Column** - Jury or televoting zmienia wartości 'J' i 'T' na przyjaźniejsze użytkownikowi biznesowemu odpowiednio 'Jury' i 'Televoting'.
— **Sort** - zapewnia odpowiednie sortowanie kolumn w hurtowni danych(po krajach, a następnie po roku konkursu) 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. Jest to serce hurtowni, którego każdy wiersz mówi o liczbie przyznanych punktów od państwa A dla państwa B w danym roku, z dodatkowymi podziałami. Oprócz tego, bezpośrednio po zczytywaniu z Internetu, skrypt dokonuje szeregu modyfikacji, generując przydatne do owocodajnej analizy miarki. Dane dotyczą tylko konkursów finałowych - zostało to postanowione ze względu na fakt, że odbywały się one zawsze i każde uczestniczące państwo było w nich uprawnione do głosowania. Przeciwnie sytuacja wygląda dla półfinałów - te niekiedy się odbywały, czasem był jeden, czasem dwa; a także i z liczbą i podziałem krajów głosujących bywało bardzo różnie; taka analiza byłaby trudna i na wielu płaszczyznach zapewne i efektująca w sprzeczne wnioski z punktu widzenia biznesowo-analitycznego. Szczegółowo, mamy tu styczność z takimi kolumnami jak:

- *ID* - unikalny identyfikator wiersza - + liczba całkowita, od 1, aż do 35349 (na ten moment)

- *Voting* - nazwa państwa głosującego, którego dotyczy dana obserwacja. Historycznie od 1975 w ramce występuje 52 różnych krajów, z czego niektóre już nieistniejące - jak chociażby Jugosławia czy Bośnia i Herzegowina. Mimo różnych nazw na przełomie lat (patrz: Macedonia) czy niespójności nazewnictwa w zależności od źródła (Bosnia Herzegovina / Bosnia and Herzegovina), finalnie po obliczeniach komputerowych jeden obszar odpowiada zawsze jednej i tej samej nazwie; od teraz skrótowo wymieniane także jako kraj / państwo A

- *Voted* - kolumna analogiczna do wspomnianej wyżej; tu jednak mamy styczność z krajami, które przytoczone głosy dostają. Uwaga! Przy analizie należy mieć na uwadze możliwie mylący przypadek z 2006. roku - Serbia Montenegro brała wtedy udział w konkursie jedynie poprzez głosowanie, lecz nie wystawiając reprezentanta [jedyny taki przypadek w historii]; od teraz skrótowo wymieniane także jako kraj / państwo B

- *Year* - rok, którego dotyczy obserwacja. Na ten moment liczba między 1975 a 2021, bez roku 2020, kiedy konkurs się nie odbył

- *Jury or televoting* - odpowiedź na pytanie: czy punkty w danym wierszu odpowiadają za te wystawione przez telewidzów, czy może zespół profesjonalnych jurorów? Uwaga! Są lata i kraje, gdzie przyznawane były jedynie punkty od ekspertów; ale są też takie, kiedy pełną moc mieli telewidzowie; oraz wreszcie takie, gdzie część jest przyznawane od jednej strony, a część od drugiej

- *Points given* - liczba punktów od państwa A dla państwa B. Zawsze jest to liczba całkowita ze zbioru 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12 - w większości przypadków 0 (każdy kraj rozdaje odpowiednie liczby punktów 10 najbardziej odpowiadających mu państwom)

- *Jury points this year* - łączna liczba punktów, którą w tym roku państwo B uzyskało od ekspertów - razem od wszystkich krajów

- *Tele points this year* - łączna liczba punktów, którą w tym roku państwo B uzyskało od telewidzów - razem od wszystkich krajów

- *All points this year* - sumaryczna liczba punktów, którą w tym roku państwo B uzyskało łącznie od jurorów i telewidzów - jest to suma dwóch wartości dwóch powyższych kolumn implikująca wraz z porównaniem z wynikami pozostałych krajów ostateczny wynik. Uwaga! Hurtownia nie zawiera informacji o sytuacjach konfliktowych, takich jak remisy punktowe krajów; w absolutnej większości przypadków wyżej w rankingu kończy wtedy kraj, który a) zdobył więcej punktów od telewidzów b) wystąpił o wcześniejszej godzinie. W przyszłości hurtownię można rozbudować o dołączenie danych zawierających informacje o ostatecznych pozycjach, które też są wolno dostępne w zasobach Internetowych

- *All points from voting to voted* - przydatna miarka; liczba wszystkich punktów, które razem państwo A przekazało państwu B, w ciągu wszystkich lat uczestnistwa. Naturalnie i tu brane są pod uwagę jedynie punkty przyznawane w finałach

- *All points from voting when voted in final* - sumaryczna liczba punktów, jaką w każdej formie przyznawało państwo A, kiedy B uczestniczyło w finale konkursu (zakwalifikowało się / było tam automatycznie / występowało, bo nie odbyły się żadne półfinały)

- *All points to voted when voting takes part* - wszystkie punkty, które uzyskało państwo B w finałach, kiedy A było uprawnione do głosowania

- *Mean countries taking part* - średnia liczba głosujących w finale w latach, kiedy występuje w nich B

- *Contests when voted in finals and voting takes part* - liczba lat, w których państwo A było uprawnione do głosowania do głosowania na B w finale

- *Sympathy coefficient* - miarka dodawana w narzędziu SAS Visual Analytics - jest to iloraz: *All points from voting to voted / Contests when voted in finals and voting takes part - All points to voted when voting takes part / (Mean countries taking part * Contests when voted in finals and voting takes part)* - mówi ona o tym, jak bardzo państwo A jest skłonne głosować na państwo B - w porównaniu do reszty świata uczestniczącej w konkursie. Wartość blisko zeru oznacza standardową chęć do głosowania (ani powyżej przeciętnej, ani poniżej); większe odchylenia od zera determinują zaś znaczną tendencję do częstszego bądź rzadszego oddawania głosów z kraju A do B - jest to naturalnie powiązane z historią, kulturą, a także geografią uczestników.

— **Jury** - Tabela wymiarów. Zawiera informacje dotyczące każdego państwa w danym roku Eurowizji (jest to złożony klucz główny Year oraz Country), dla konkretnych lat - informacje uzależnione od aktualnego stanu w Internetowym serwisie *data.world*. Oryginalny zbiór został mocno okrojony, przetwarzając szczegółowe informacje o jurorach na zbiorze informacje w formie kategorii. Wśród kolumn znajdują się:

- *Year* - rok, którego dotyczy obserwacja. W przypadku braku informacji, brakujące lata i kraje uzupełniane są wartościami *Unknown*

- *Country of jury* - kraj, którego dotyczą niżej opisane informacje o składzie eksperckim

- *Male count* - liczba mężczyzn, która w danym roku i kraju zasiadła w jury. Cały skład zawsze liczy 5 osób, więc jest to wartość całkowita między 0 a 5

- *Mean age category* - zkategoryzowana średnia wieku składu jurorskiego. Na podstawie historycznego rozkładu, przygotowane zostało grupowanie:
— Wiek 35 lat lub mniej: *Very young*; jedna z dwóch najmniej licznych grup
— (35; 38] lat: *young*; liczna grupa
— (38; 43] lat: *middle*; najliczniejsza grupa
— (43; 50] lat: *old*; liczna grupa
— Wiek ponad 50 lat: *Very old*; jedna z dwóch najmniej licznych grup
- *Female count* - kategoria utworzona na podstawie miarkiw procesie ETL - liczba kobiet, wyliczona na podstawie faktu o stałości liczności składu jurorskiego oraz kolumny *Male count*

— **GeographicDimension** - Tabela wymiarów. Hierarchia geograficzna. Ramka została przygotowana ręcznie i przy sytuacjach konfliktowych została podjęta odpowiednia analiza ku trafnemu przypisaniu odpowiednich doszczegółowień. Każdemu państwu biorącemu udział w Konkursie Piosenki Eurowizji (klucz główny) kolumny doszczegółowujące przestrzenne informacje o krajach:

- *Country* - nazwa państwa; rozpatrywane wszystkie odpowiednio połączone z tabelą faktów

- *Continent part* - część kontynentu, do której można zaliczyć kraj. Łącznie uznanych zostało 11 kategorii dla 52 państw: Australia, Central Europe, Eastern Europe, North Africa, Northern Europe, Northwestern Europe, Southeast Europe, Southern Europe, Southwestern Europe, Western Asia i Western Europe

Continent - informacja o jednym z czterech kontynentów, do którego należy kraj

— **TimeDimension** - Tabela wymiarów. Wymiar Czasowy. Jego trzy kolumny obejmują:

- *Year* - rok

- *Decade* - dekada

- *century* - stulecie

— **Contestants** - Tabela wymiarów. Jej kluczem głównym jest rok i kraj - dla każdej pary doszczegółowione są informacje o uczestniku. Ze względu na analizowane lata od 1975. naturalnie każdemu roku przypada tylko jedna piosenka, a co za tym idzie jedne dane określające ją. Brakujące wartości zostały zastąpione łańcuchem znaków *Unknown*. Ramka została przetworzona z oryginalnej *contestants.csv* (dostępnej także w katalogu */data-/original*) z Internetu ku poprawionej do celów biznesowo analitycznych, poprzez skrypt *original-contestants-to-dim-table.R*. Wyróżnione zostały takie kolumny jak:

- *Year* - rok, w którym zgłoszona została propozycja muzyczna

- *Country* - kraj, który reprezentuje piosenka

- *Performer* - nazwa artysty

- *Song* - tytuł piosenki

- *Composers* - kompozytor(zy) utworu

- *Lyricists* - autor(zy) tekstu propozycji

- *Lyrics length category* - zkategoryzowana informacja o długości tekstu piosenki. Zastosowane zostało grupowanie:

— Co najwyżej 750 znaków: *Very short*
— (750; 1000) znaków: *Short*
— (1000; 1250¿ znaków: *Middle*
— (1250; 1500¿ znaków: *Long*
— Powyżej 1500 znaków: *Very long*

— **GDP** - Tabela wymiarów. Zawiera ona kolumny dotyczące krajowych produktów brutto analizowanych państw. Na ramkę składają się konkretnie:

- *Country* - brane pod uwagę państwo

- *Year* - rok, którego dotyczy wiersz danych

- *Current US digits* - liczba cyfr produktu brutto w dolarach

- *Growth annual* - informacja o tym, jak zmienił się krajowy produkt brutto w stosunku do roku poprzedniego. Przyjęte zostało następujące grupowanie:

— Spadek +10%: *Huge decrease*
— Spadek +5%, ale poniżej 10%: *Big decrease*
— Spadek +1%, ale poniżej 5%: *Slight decrease*
— Wahania między -1% a +1%: *Stagnation*
— Wzrost +1%, ale poniżej 5%: *Slight increase*
— Wzrost +5%, ale poniżej 10%: *Big increase*
— Wzrost +10%: *Huge increase*

W wypadku braków danych przyjmowane są wartości *Unknown*

Uwaga! Tabela faktów jest typu *periodic snapshot* - konkurs odbywa się co roku w maju i wtedy też aktualizowane są dane w serwisie *data.world*. Dlatego też docelowo zczytywanie i przetwarzanie *Points given*, które w wersji kompletnej trwa kilkanaście godzin, powinno być wykonywane **najrzadziej raz do roku**. Można rozważyć także częstsze pobieranie informacji - w tym momencie oryginalna tabela scrapingowana z Internetu zawiera kilka błędów, takich jak niespójność niektórych nazw państw czy pojedyncze pomylone wartości, takie jak zamiana Litwy z Łotwą. Skrypt kontroluje owe turbulencje i je poprawia, natomiast trzeba mieć na uwadze, że mogą one zostać poprawione niekoniecznie w sezonie konkursowym.

6. Warstwa raportowa

6.1. Opis warstwy raportowej

Warstwa raportowa wykorzystuje system SAS Visual Analytics - korzystaliśmy z SAS Trial Viya. Posłużyła ona do wykonania ciekawych wizualizacji i raportów zawierających godne uwagi dane uzyskane z pomocą hurtowni i procesu ETL.

6.1.1. Proces analizy danych

Cały proces tworzenia analizy danych został zrealizowany w czterech krokach:

1. Model gwiazdy został zaimportowany do SAS Trial Viya w ramach dostępnego modułu **Zarządzaj danymi** jako lokalny plik .csv.

2. W module **Eksploracja i wizualizacja** stworzono miary kalkulowane, miary zagregowane oraz hierarchie

3. W powyższym module również stworzono strony, na których powstały wizualizacje - w zależności od sposobu wizualizacji danych, wykorzystano różne miarki wraz z kategoriami lub hierarchiami.

4. Odpowiednie wykresy filtrowano w zależności od użytych danych w celu otrzymania zamierzonych wyników odpowiadających np. danemu okresowi czasu albo danemu państwu.

6.1.2. Hierarchie w części raportowej

W warstwie raportowej stworzono poniższe 3 hierarchie:

— **Time hierarchy** : Century -¿ Decade - Year
— **Voting hierarchy** : Continent - Continent part - Voting
— **Voted hierarchy** : Continent - Continent part - Voted

Oddzielne hierarchie dla Voting i Voted (country) zostały stworzone ze względu na fakt, iż głosujących krajów było 52, a tych na które głosowano było 51.

6.1.3. Transformacje w części raportowej

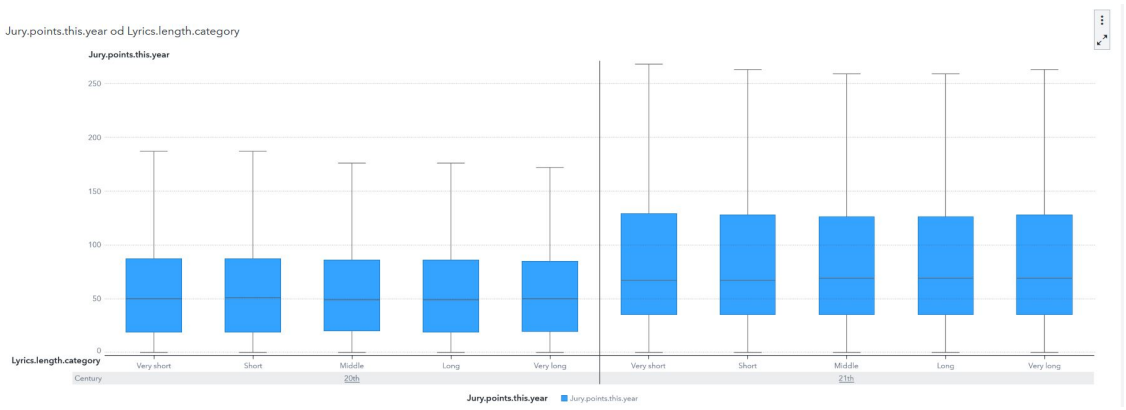
Użyte narzędzie powalało na automatyczny dobór typu agregacji miarek przy dodawaniu ich do danych wizualizacji. Utworzono również nową miarkę kalkulowaną **Sympathy coefficient**, transformując przy tym już istniejące. Poniżej przedstawiono schemat tworzenia miarki Sympathy coefficient - współczynnika sympatii:

((All.points.from.voting.to.voted / Contests.when.voted.in.finals.and.voting.takes.part) - (All.points.to.voted.when.voting.takes.part / (Mean.countries.taking.part * Contests.when.voted.in.finals.and.voting.takes.part)))

Rys. 6.1. Miara kalkulowana - współczynnik sympatii

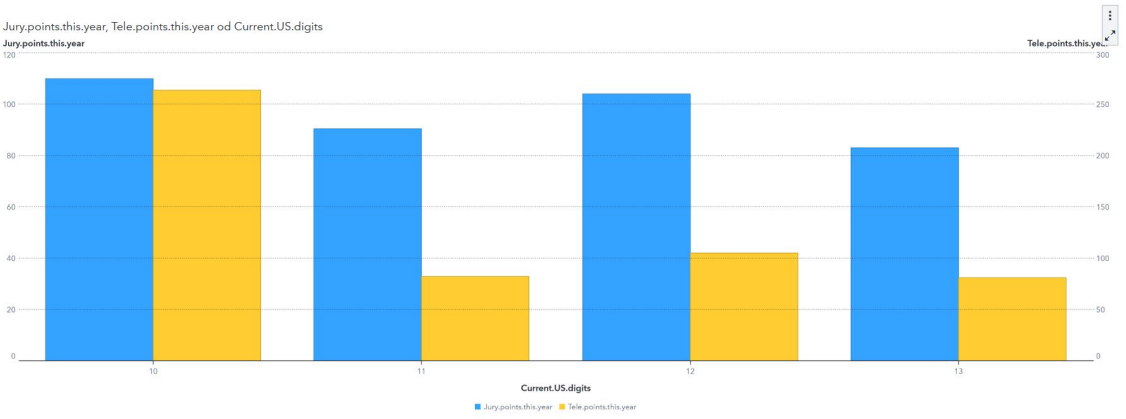
6.2. Przykładowe raporty dla użytkownika

Możliwości narzędzia jakim jest SAS pozwoliły nam na stworzenie poniższych wizualizacji:
- porównanie wpływu długości piosenek na otrzymywane punkty od Jury(wzrost w XXI wieku wynikał ze wzrostu liczby państw biorących udział, co przełożyło się na zwiększenie średnio otrzymywanej liczby punktów)



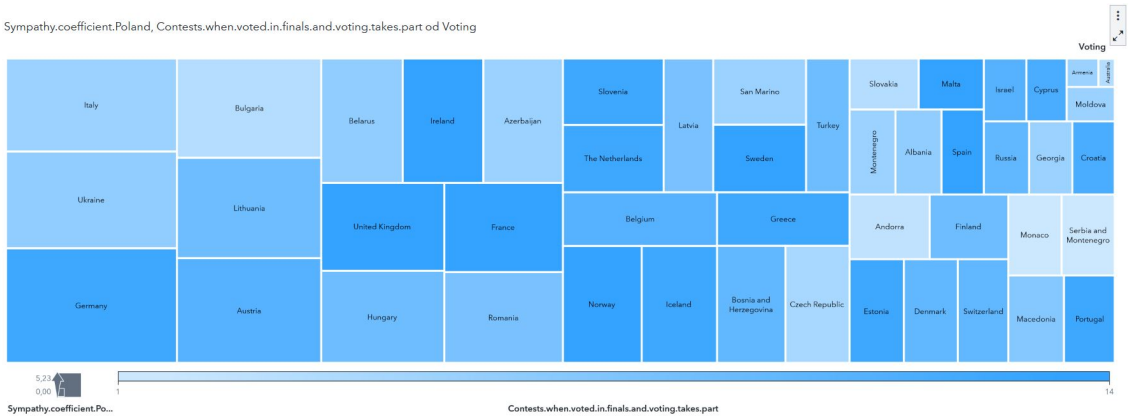
Rys. 6.2. Wpływ długości piosenek na punkty otrzymywane od jury

- wizualizacja dotycząca zależności wielkości PKB danego państwa (liczba cyfr), a otrzymywanej przez niego punktacji od jury oraz od telewidzów



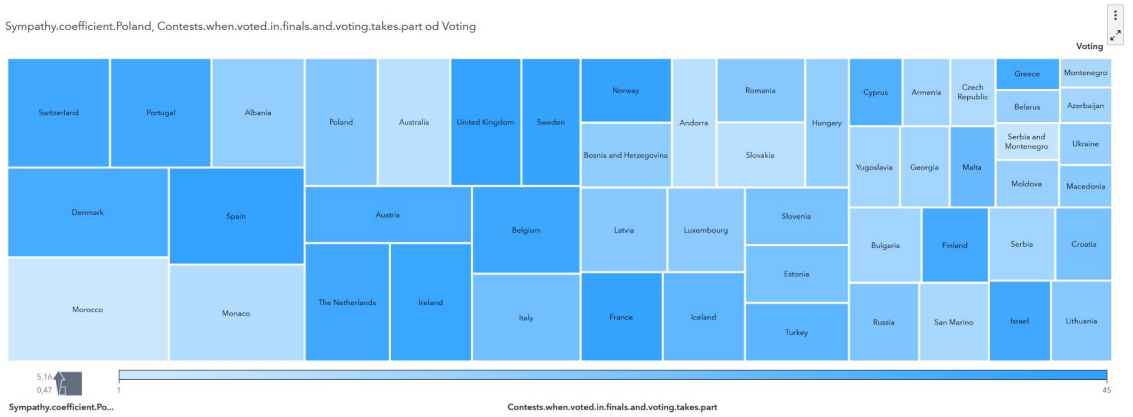
Rys. 6.3. Wpływ zamożności państw (PKB) na uzyskiwane punkty

- wizualizacja do porównywania współczynników sympatii od danych państw dla Polski



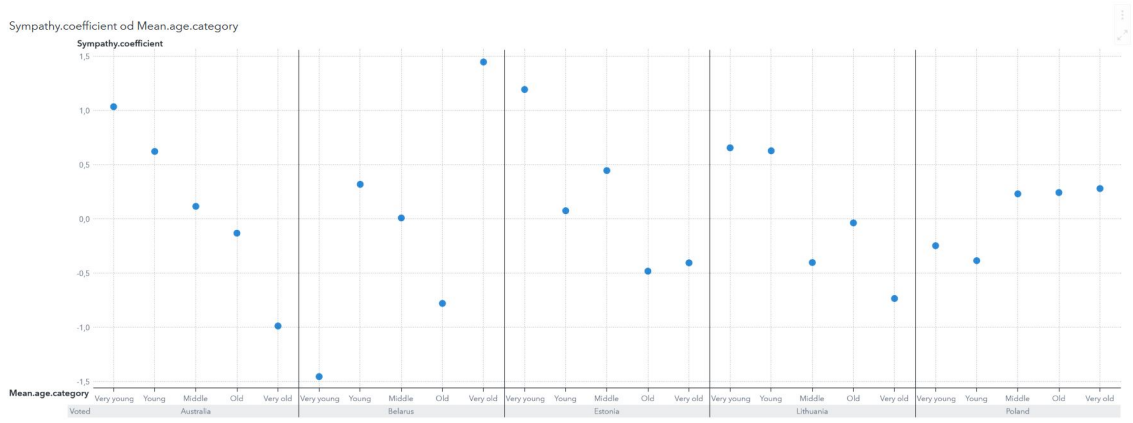
Rys. 6.4. Mapa ciepła - współczynniki sympatii dla Polski

- wizualizacja do porównywania współczynników sympatii od danych państw dla Niemiec (porównawczo z Polską)



Rys. 6.5. Mapa ciepła - współczynniki sympatii dla Niemiec

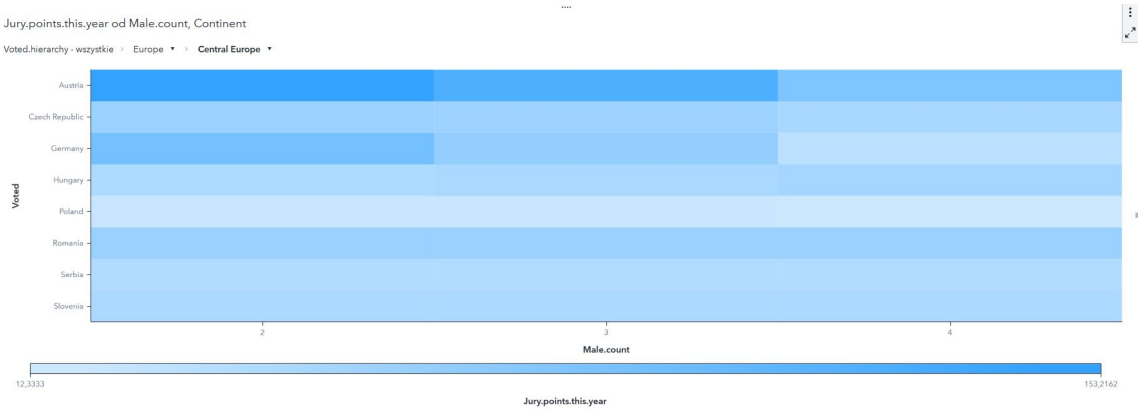
- wizualizacja badająca związek wieku jury ze współczynnikami sympatii dla danego państwa



Rys. 6.6. Współczynniki sympatii w zależności średniego wieku jury dla wybranych państw

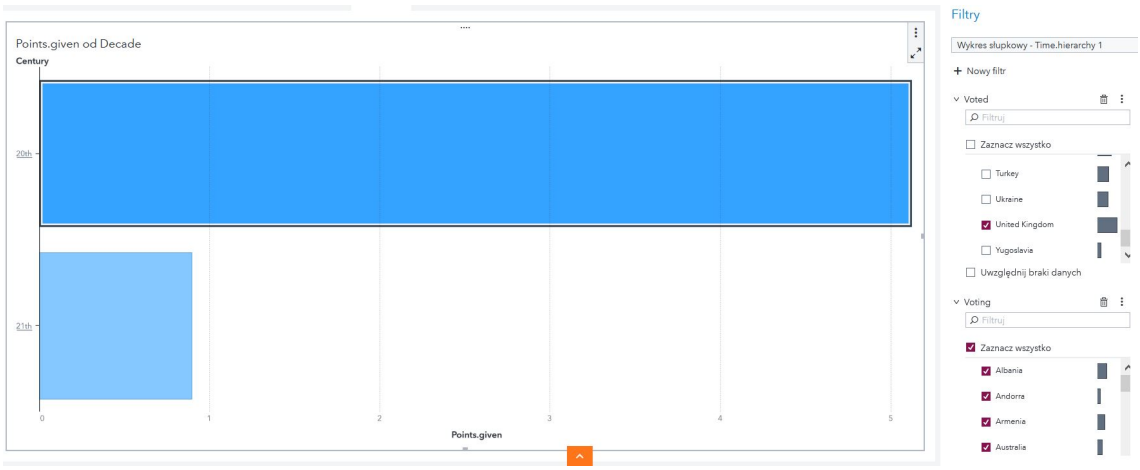
- wizualizacja dotycząca punktacji od jury badająca wpływ składu płci na uzyskane wyniki

od jury na tle hierarchii geograficznej



Rys. 6.7. Punkty od jury w zależności od składu płci w jury na tle hierarchii geograficznej

- wizualizacja dotycząca punktacji indywidualnego państwa (Wielkiej Brytanii) na tle hierarchii czasowej Time hierarchy



Rys. 6.8. Punkty uzyskane przez Wielką Brytanię w XX oraz XXI stuleciu

7. Wyniki projektu

Trudno jednoznacznie wskazać, czy ostateczne rezultaty są w pełni satysfakcjonujące. Z jednej strony swego rodzaju celem było wskazanie, że nie ma jakiejś magicznej recepty na wygranie Konkursu Piosenki Eurowizji, jednak z drugiej po cichu miało się nadzieję na pewne nieoczekiwane wnioski.

Z pewnością faktem jest, że polityka, historia oraz geografia mają duże znaczenia - państwa leżące na mapie koło siebie rzeczywiście częściej mają tendencję do obdarowywania się większą niż typowa liczba punktów (patrz zwłaszcza: Grecja i Cypr); są też jednak przypadki zupełnie odwrotnego efektu - tak jak dla Armenii i Azerbejdżanu. Nie jest to jednak główny wyznacznik sukcesu, i "mniej popularne" państwa również mają szansę na dobry wynik bądź zwycięstwo - choć co pokazują dane, zdecydowanie niektórym krajom jest ciężej o takie rezultaty.

Na pewno naszą dumą jest opracowanie tworzonego przez nas w procesie ETL oraz w SAS VA współczynnika sympatii - niejawna miara bardzo dobrze odzwierciedla realia morale krajów i ma obiektywnie ogromną wartość z punktu widzenia biznesowo-analitycznego.

Co ciekawe, nie zaobserwowano istotnych zależności w kwestii głosowania jurorów i telewidzów na utwory dłuższe czy krótsze pod względem lirycznym - co okazało się być sprzeczną z postawioną przez nas tezą, że zapewne telewidzowie prędzej upodobują sobie utwory prostsze tekstowo, zaś eksperci - te złożone.

Niezbadana do końca pozostaje sprawa głosów jurorów w zależności od ich demografii - co prawda nasza warstwa raportowa wykazała pewne istotne własności, natomiast nie skorzystaliśmy jednak obiektywnie z wystarczająco reprezentatywnej próbki. Z dużym prawdopodobieństwem dane państwa częściej wystawiają ekspertów o podobnym bilansie płci czy średnim wieku, co zdecydowanie zaburza wnioskowanie na tym etapie. Przeszkodą okazała się także stosunkowo duża liczba braków w temacie danych o jurorach.

Temat nie jest jednak jeszcze wyczerpany i "kod Eurowizji" jest potencjalnie do odkrycia. W analizie nie były chociażby brane pod uwagę pozycje startowe i porównywanie ostatecznych wyników rankingowo (a nie jedynie przyglądając się liczbie punktów), co z pewnością przyniosłoby pewną ilość interesujących wniosków i faktów wynikających z historycznych edycji. Prawdą jest chociażby, że zazwyczaj w finale średnio radzi sobie kraj występujący jako drugi, zaś zaskakująco dobrze państwa prezentujące swój utwór pod koniec. Mimo, że pozycja w większości przypadków była narzucana z góry przez organizatorów, prawdopodobnie przyjrzenie się tematowi z tej perspektywy również dało by ciekawe rezultaty.

Podsumowując - budując hurtownię danych osiągnęliśmy sukces w analizie Konkursu Piosenki Eurowizji biorąc pod uwagę docelowo użyte informacje, natomiast w przypadku posiadania większej ilości czasu efekt mógłby być jeszcze lepszy. Abstrahując od samej wartości analizy i procesu - z pewnością był to też dobry projekt rozwojowo i pozwolił on dobrze zrozumieć procesy ETL, a także narzędzia związane z raportowaniem hurtowni danych ;)

8. Testy funkcjonalne

Ku kontroli poprawności procesu ETL zostało przygotowanych kilka testów. Są to proste skrypty SQL modyfikujące konkretne wartości w wybranych tabelach. Rozpatrzone zostały różne przypadki, które mogły potencjalnie zaburzyć pracę oprogramowania i poprawnie przeszły one zaproponowane kryteria. Niestety, rozwiązanie nie jest odporne na wszelkie możliwe, acz mało prawdopodobne przypadki - faktem jest chociażby typ danych długości państw, który zawsze przyjmuje maksymalnie dokładnie 22 znaki - tyle, ile ma "Bosnia and Herzegovina". W przypadku debiutu kraju o dłuższej nazwie należałoby przyjrzeć się poszczególnym komponentom i dostosować je do nowej sytuacji. To samo dotyczy się ręcznego uzupełnienia hierarchii geograficznej. Średnio co kilka lat zmieniane są także nieco zasady głosowania - dokonaliśmy wszelkich starań, aby nasz projekt niezależnie od potencjalnych form był także dostosowany do możliwych scenariuszy w przyszłości; niewykluczone jest jednak, że w przyszłości także należałoby zastanowić się i zrealizować nowy sposób przetwarzania danych - taka sytuacja mogłaby się wydarzyć chociażby w teoretycznej sytuacji wprowadzenia trzeciego bloku przyznawania punktów - głosowania internetowego; bądź w dyskutowanej koncepcji zmiany proporcji wagi głosów od jurorów i telewidzów (zawsze było to 50:50, lecz w przyszłości możliwe są zmiany).

Przytoczone testy znajdują się także w katalogu */tests* - należy je uruchamiać w SQL Server Management Studio i obserwować zachowanie Visual Studio Integration Services.

8.1. Podmiana informacji o przyznanych punktach w tabeli faktów

```
---Edit

UPDATE [Points Given]
SET [Points given] = 12,
[All points from voting to voted] = 13,
[All points to voted when voting takes part] = 1572
WHERE Year = 2017 and Voting = 'Armenia' AND Voted = 'Azerbaijan'
    AND [Jury or Televoting] = 'Televoting';

UPDATE [Points Given]
SET [Points given] = 0, ---[Tele points this year] = ,
[All points from voting to voted] = 38,
[All points to voted when voting takes part] = 735
WHERE Year = 2017 and Voting = 'Armenia' AND Voted = 'Cyprus'
    AND [Jury or Televoting] = 'Televoting';

---Back to true values

UPDATE [Points Given]
SET [Points given] = 0,
[All points from voting to voted] = 1,
[All points to voted when voting takes part] = 1560
WHERE Year = 2017 and Voting = 'Armenia' AND Voted = 'Azerbaijan'
```

```
    AND [Jury or Televoting] = 'Televoting';

UPDATE [Points Given]
SET [Points given] = 12, ---[Tele points this year] = ,
[All points from voting to voted] = 50,
[All points to voted when voting takes part] = 747
WHERE Year = 2017 and Voting = 'Armenia' AND Voted = 'Cyprus'
    AND [Jury or Televoting] = 'Televoting';
```

8.2. Uzupełnienie danych o jurorach

```
---Edit

UPDATE [Jury]
SET [Male count] = '3', [Female count] = '2',
[Mean age category] = 'Young'
WHERE Year = 2021 and [Country of jury] = 'Poland'

---Back to true values

UPDATE [Jury]
SET [Male count] = 'Unknown', [Female count] = 'Unknown',
[Mean age category] = 'Unknown'
WHERE Year = 2021 and [Country of jury] = 'Poland'
```

8.3. Modyfikacja wymiaru geograficznego

```
---Edit

UPDATE [Geographic Dimension]
SET [Continent part] = 'Southeast Europe',
[Continent] = 'Europe'
WHERE Country = 'Cyprus'

---Back to true values
UPDATE [Geographic Dimension]
SET [Continent part] = 'Western Asia',
[Continent] = 'Asia'
WHERE Country = 'Cyprus'
```

Proces ETL poprawnie przetwarza wszystkie trzy testy.

9. Podział pracy

Jakub Kosterna - model hurtowni danych, testy funkcjonalne
Patryk Wrona - ETL, SAS

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