Turystyka w UE

2023

—

Maciej Orsłowski, Sabina Sidarovich

15

Cel projektu oraz planowane korzyści dla odbiorcy

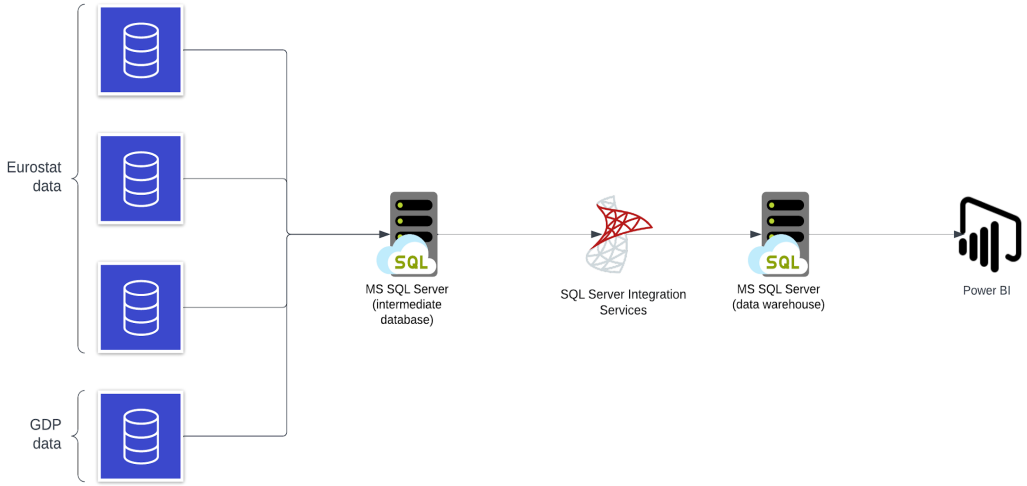
Celem projektu jest stworzenie interaktywnych analitycznych raportów, które umożliwią użytkownikom przegląd ogólnych trendów turystycznych wśród obywateli Unii Europejskiej (UE) oraz dostarczą cenne informacje dotyczące wydatków, wzorców podróży i danych demograficznych.Narzędzie ma umożliwiać analizę i zrozumienie turystyki w UE i poza nią. Dashboardy będą oferować interaktywne funkcje, które pozwolą użytkownikom na bardziej szczegółową analizę danych i głębsze zrozumienie trendów i wzorców w turystyce.

Planowane korzyści dla odbiorcy rozwiązania:

- Szczegółowa analiza trendów turystycznych: Odbiorca będzie miał możliwość dokładnego zbadania trendów turystycznych w UE i poza nią. Dashboardy umożliwią identyfikację kluczowych wzorców podróży oraz wydatków z nimi związanych.
- Benchmarking i porównania: Rozwiązanie umożliwi odbiorcy porównanie danych turystycznych między różnymi krajami UE oraz porównanie wyników UE z danymi spoza UE. To pozwoli na ocenę konkurencyjności turystycznej, identyfikację mocnych i słabych stron oraz potencjalnych obszarów wzrostu.
- Prognozowanie i podejmowanie decyzji: Przedstawiona analityka umożliwi odbiorcy podejmowanie lepiej poinformowanych decyzji biznesowych.
- Źródło wiarygodnych danych: Rozwiązanie będzie oparte na sprawdzonych danych z Eurostatu oraz innych wiarygodnych źródeł. Odbiorca będzie miał pewność, że korzysta z aktualnych, wiarygodnych i kompleksowych danych turystycznych.

Dzięki tym korzyściom, odbiorca rozwiązania będzie w stanie lepiej zrozumieć i analizować trendy turystyczne, podejmować bardziej informowane decyzje biznesowe oraz doskonalić swoje strategie marketingowe w kontekście turystyki w UE i poza nią.

Narzędzia

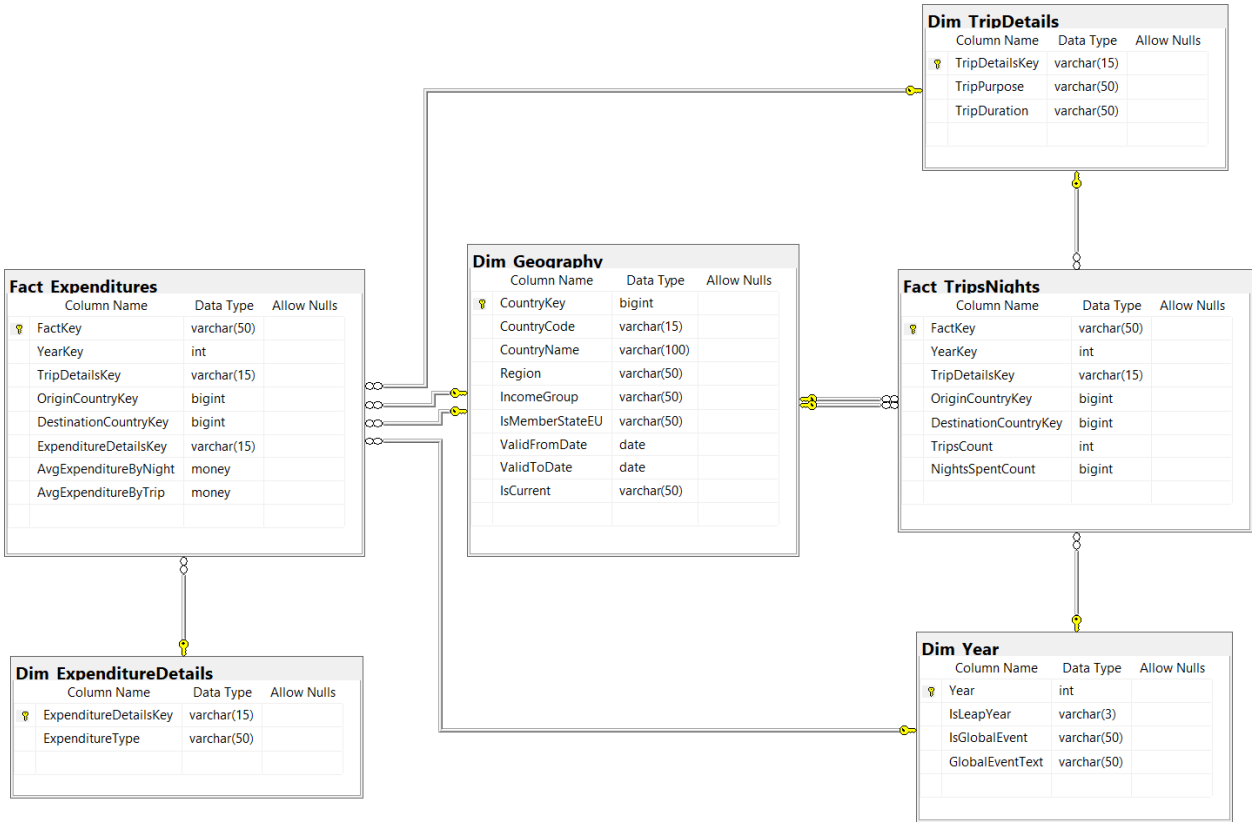


Architektura wykorzystuje MS SQL Server do przechowywania i zarządzania danymi, MS SSIS do procesów ETL oraz Power BI do wizualizacji danych i tworzenia dashboardów. Ta kombinacja narzędzi zapewnia solidne kompleksowe rozwiązanie do przygotowywania i prezentowania dashboardu z pożądanymi analizami i spostrzeżeniami.

Źródła danych

Nazwa	Link	Ostatnia aktualizacja	Stan	Format
Nights spent by duration, purpose and main destination of the trip	ec.europa.eu/eurostat/web/products-datasets/-/tour_dem_tnw	12.05.2023 (odświeżane rocznie)	active	csv
Expenditure by duration, purpose, main destination of the trip and expenditure category	ec.europa.eu/eurostat/web/products-datasets/-/tour_dem_extotw	12.05.2023 (odświeżane rocznie)	active	csv
Trips by duration, purpose and main destination	ec.europa.eu/eurostat/web/products-datasets/-/tour_dem_ttw	12.05.2023 (odświeżane rocznie)	active	csv
Country GDP	data.worldbank.org/indicator/NY.GDP.PCAP.CD	(odświeżane rocznie)	active	csv

Model Hurtowni



W kontekście wymagalności danych, żadna z kolumn nie akceptuje wartości NULL, z wyjątkiem ValidToDate (SCD2), gdzie brak wartości wskazuje na aktualność danych. Rekordy z brakującymi danymi w miarkach są usuwane w trakcie procesu ładowania faktów w SSIS. W przypadku pozostałych kolumn, takich jak klucze obce do wymiarów, ze względu na ich niedopuszczalność wartości NULL, ewentualne braki danych zostaną wykryte w SSIS jako błąd podczas próby umieszczenia danych w hurtowni. Pozwala to na inspekcję tych danych, jednocześnie zapewniając, że rekordy z brakującymi danymi nie zostaną umieszczone w hurtowni.

Opis procesu ETL:

1. Pobieranie i ładowanie danych (package Scraping.dtsx)

    1.1. Dane pobierane są ze źródeł za pomocą zapytań http do wystawionych przez dostawców danych endpointów. Są one następnie zrzucane do plików csv. Całość odbywa się za pomocą skryptu napisanego w języku Python.

    1.2. Pliki csv są następnie wczytywane do pośredniej bazy danych, której tabele odzwierciedlają strukturę tychże plików. Krok ten ma za zadanie po pierwsze uprościć data flow podczas wczytywania danych do hurtowni (dane wczytywane w SSIS z plików płaskich domyślnie mają typ varchar(50), konieczne jest więc mapowanie np. kolumn liczbowych, które to właśnie odbywa się już na tym etapie). Po drugie, takie działanie ułatwia potem testowanie ETL, gdyż na pośredniej bazie można wywoływać polecenia SQL, czego nie można szybko zrobić na pliku płaskim.

2. Wczytywanie wymiarów (package LoadDimensions.dtsx, poza Dim_Year które jest wypełniane za pomocą skryptu Dim_Year.sql)

    2.1. **Dim_TripDetails**:

    - ***TripDetailsKey*** - sklejone *purpose* oraz *duration* z tabel źródłowych
    - ***TripPurpose***, ***TripDuration*** - znajdujemy wszystkie występujące kombinacje kolumn *purpose* i *duration* z tabel *expenditures*, *trips* oraz *nights* i umieszczamy w tym wymiarze.

      Pomijane są zagregowane kategorie, czyli rekordy z wartością “TOTAL” w kolumnie *purpose* (bo wartości tych rekordów są sumą analogicznych rekordów z kategoriami “PER” oraz “PROF”), a także rekordy z warością “N_GE1” w kolumnie *duration* (kod oznacza “1 lub więcej noc” i analogicznie jak wcześniej jest to suma rekordów z “N1-3” oraz “N_GE4”).

      Usunięta jest też kategoria celu podróży “PER_VFR” (personal visits to family or relatives), ze względu na liczne braki w danych. Nie prowadzi to do utraty danych gdyż tego typu podróże są wliczone w kategorię personal.

      Pozostałe wartości są mapowane na sensownie brzmiące nazwy w następujący sposób:

Kolumna	Wartość	Mapowanie
purpose	PER	Personal reasons
purpose	PROF	Professional, business
duration	N1-3	From 1 to 3 nights
duration	N_GE4	4 nights or over

2.2. **Dim_ExpenditureDetails**:

- ***ExpenditureDetailsKey*** - *expend* z tabeli źródłowej
- ***ExpenditureType*** - kolumna zawierająca wszystkie unikalne wartości z kolumny *expend* z tabeli źródłowej *expenditures*, zmapowane na sensownie brzmiące nazwy w następujący sposób:

Wartość	Mapowanie
TRA	Transport
REST	Restaurants/cafes
ACCOM	Accomodation
N_GE4	4 nights or over
DUR	Durables and valuable goods
PACK_ARR	Package arrangement
TRP_OTH	Other expenditure

Usunięta została jedynie kategoria “TOTXDUR” dla której wartości faktów to suma rekordów dla pozostałych kategorii z wyjątkiem “DUR”.

2.3. **Dim_Geography**:

- Tabelą źródłową są metadane o krajach z danych o GDP.
- ***CountryKey*** - identity column, sztuczny klucz główny stworzony na potrzeby SCD2
- ***CountryCode*** - 2-znakowy kod kraju ISO z dwoma wyjątkami: Grecja, która ma kod EL oraz Wielka Brytania, która ma kod UK. Wyjątki są spowodowane tym, że EU wykorzystuje dla tych krajów inne kody niż

te definiowane przez ISO, stąd potrzeba ich zmapowania w celu zgodności z danymi z Eurostatu.

- Kolumny ***Region*** oraz ***IncomeGroup*** są niezmienione względem źródła.
- ***CountryName*** - kolumna *TableName* ze źródłowej tabeli.
- ***IsMemberStateEU*** - powstaje poprzez mapowanie kodu kraju na wartość logiczną informującą czy dany kraj przynależy do Unii Europejskiej czy nie. Przy pierwotnym wczytywaniu danych do tego wymiaru należy uwzględnić dotychczasowe zmiany członkowskie od roku 2012 (początek danych)
- ***ValidFromDate***, ***ValidToDate***, ***IsCurrent*** - od kiedy do kiedy i czy dane są aktualne (SCD typu 2)
- SCD2 śledzi zmiany we wszyskich kolumnach oprócz CountryCode, które uznawane jest za wartość stałą i jednoznacznie określającą dany kraj.
- Do tej tabeli doklejane są też ręcznie stworzone rekordy odpowiadające pewnym grupom krajów, dane dotyczące których zostały zgrupowane w danych z Eurostatu (np. “Other African countries”). Tak doklejone rekordy mają przydzielone sensowne wartości w kolumnie Region jeśli jest to możliwe (w przeciwnym wypadku “Not applicable”), natomiast w kolumnie IncomeGroup umieszczana jest informacja “Not applicable”.

2.4. **Dim_Year**:

- Generowane za pomocą skryptu na zapas od roku 2012 (rok początku danych). Każdy rekord odpowiada jednemu rokowi (co wynika z ziarnistości wykorzystywanych danych).
- ***Year*** - klucz główny, jest to po prostu rok
- ***IsLeapYear*** - czy rok jest przestępny
- ***IsGlobalEvent***, ***GlobalEventText*** - kolumny informujące o istotnych wydarzeniach w skali świata w danym roku, które mogą mieć wpływ na turystykę. Dobrym przykładem jest pandemia COVID w latach 2020 i 2021, wtedy te kolumny mają wartości odpowiednio “Yes” i “COVID-19 Pandemic”. Jeśli w danym roku nie miało miejsce żadne tego typu wydarzenie, to domyślne wartości to odpowiednio “No”, “No major event”. Obecność tych kolumn w wymiarze dat oznacza, że w

przypadku wystąpienia podobnych wydarzeń w danym roku wymiar powinien być odpowiednio zaaktualizowany.

3. Wczytywanie faktów (package LoadFacts.dtsx):

3.1. **Fact_Expenditures**:

- Miarki: ***AvgExpenditureByNight*** oraz ***AvgExpenditureByTrip*** są generowane za pomocą kolumn *statinfo* i *OBS_VALUE* z tabeli źródłowej *expenditures*. Wykorzystywana jest tu operacja pivot. *OBS_VALUE* zawiera wartość wydatków w €, natomiast *statinfo* zawiera informacje o sposobie agregacji (łączne / uśrednione na 1 noc / uśrednione na 1 wycieczkę). Rekordy z brakami wartości w tych miarkach są pomijane.

- Zdecydowaliśmy się zrezygnować z danych o sumarycznej wartości wydatków (czyli gdy statinfo = "TOTAL") z uwagi na to że były one praktycznie niedostępne na takim poziomie ziarnistości jak te uśrednione wartości wydatków (było dosłownie tylko ok. 1000 rekordów i wszystkie dotyczące tylko jednej z kategorii wydatków). Stąd, aby w raportach wykorzystać sumaryczną wartość wydatków, już w samym narzędziu BI został zrobiony inner join z drugą tabelą faktów i pomnożona została średnia wartość wydatków na noc przez liczbę nocy.

- ***FactKey*** - sklejenie wartości z wykorzystywanych kolumn kategorycznych z tabeli źródłowej w celu jednoznacznej identyfikacji faktu (stworzenie takiego klucza znacząco ułatwia potem ewentualną aktualizację wartości faktu oraz upraszcza skrypt do dodawania faktów)

- ***OriginCountryKey*** - kraj pochodzenia osoby udającej się na wycieczkę

- ***DestinationCountryKey*** - kraj będący głównym celem wycieczki

3.2. **Fact_TripsNights**:

- Miarki ***TripsCount*** i ***NightsSpentCount*** to wartości z kolumny *OBS_VALUE* odpowiednio z tabel *trips* i *nights*. W przypadku braku danych w jednej z tych miarek rekord jest pomijany.

Warstwa raportowa

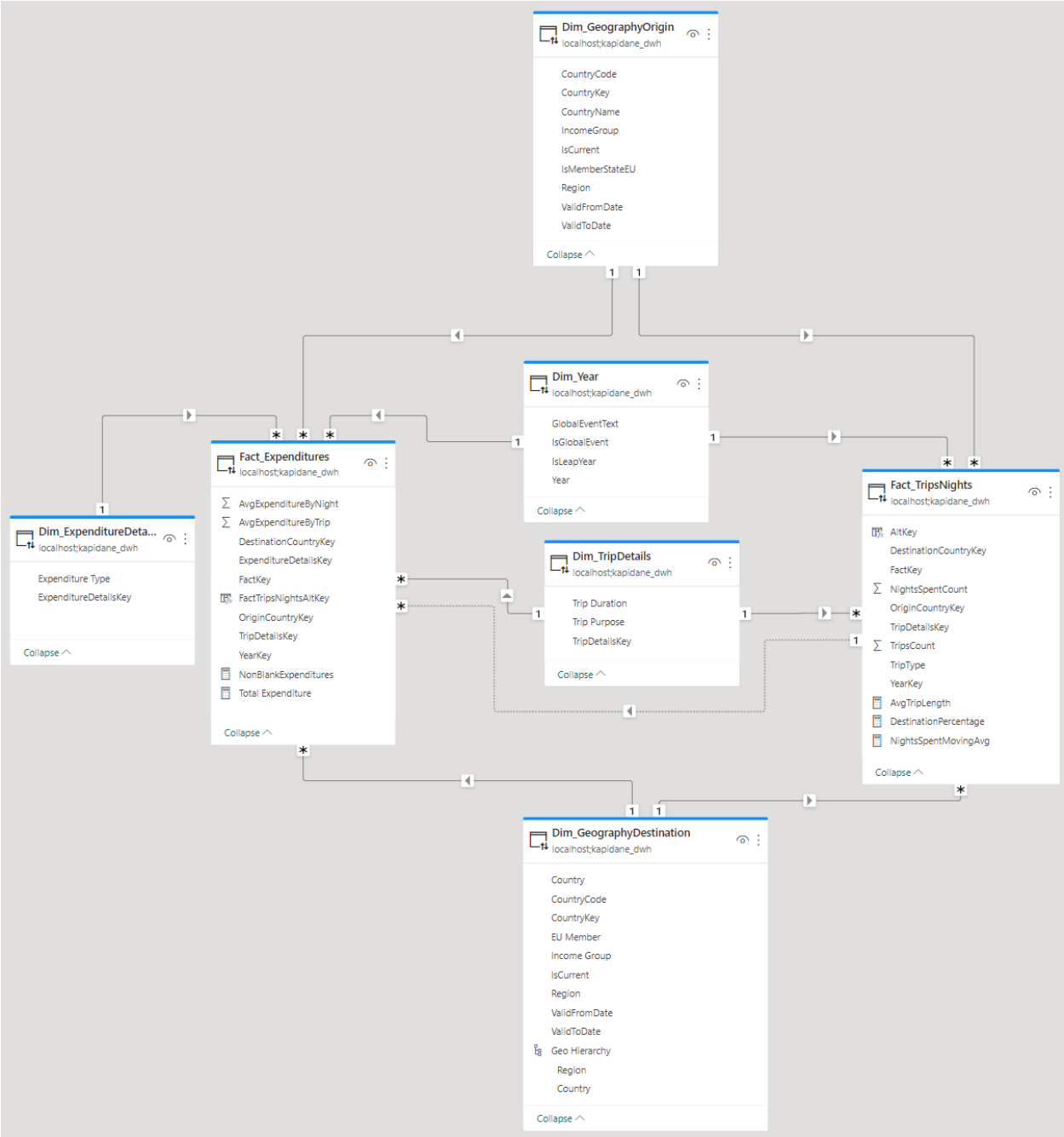
Warstwa raportowa zapewnia użytkownikom dostęp do danych zgromadzonych w hurtowni danych poprzez interfejsy wizualne w postaci przygotowanych raportów. Obejmuje ona zarówno modele biznesowe danych, jak i hierarchie oraz transformacje, które umożliwiają użytkownikom analizę i prezentację informacji w przystępny sposób.

Dane są pobierane z lokalnego serwera MS SQL Server za pomocą bezpośredniego zapytania (Direct Query) i łączone z narzędziem Power BI, umożliwiając dynamiczny dostęp do aktualnych informacji oraz zoptymalizowaną wydajność.

Warstwa opiera się na modelu biznesowym, który jest stworzony w oparciu o wymagania biznesowe i strukturę hurtowni danych. W celu umożliwienia zaawansowanej analizy i wizualizacji danych, struktura hurtowni została rozszerzona o duplikat tabeli wymiarów "Dim_Geography". Dodatkowo, została utworzona hierarchia geograficzna zawierająca kolumny "Region" i "CountryName". Obliczane są również miarki, które są istotne dla analizy poszczególnych aspektów tematycznych:

- AvgTripLength (Facts_TripsNights) - średnia długość wycieczek, wyliczana jako iloraz miarek NightsSpentCount i TripsCount
- AltKey (Facts_TripsNights), FactsTripNightsAltKey (Fact_Expenditures) - klucze obce powstałe poprzez sklejenie wartości z kolumn YearKey, OriginCountryKey, DestinationCountryKey i TripDetailsKey. Pozwalają dodać połączenie w modelu między tabelami faktów które jest konieczne do stworzenia miarki TotalExpenditure.
- DestinationPercentage (Facts_TripsNights) - Ile procent wycieczek z konkretnego kraju i dla zadanych filtrów odbywa się do danego kraju docelowego.
- NightsSpentMovingAvg (Facts_TripsNights): ruchoma średnia wartości miarki NightsSpentCount. Dla danego roku przyjmuje ona średnią wartość z trzech ostatnich lat.
- Trip Type (Facts_TripsNights): przyjmuje jedną z dwóch wartości, “Domestic” lub “International” i jest obliczana na podstawie wyniku relacji równości kluczy OriginKey i DestinationKey faktu.
- Total Expenditure (Fact_Expenditures) - Łączna liczba wydatków, tworzona poprzez pomnożenie miarek AvgExpenditureByNight i NightsSpentCount.
- NonBlankExpenditures (Fact_Expenditures): miarka pomocnicza potrzebna do tego, by filtr “Destination Country” wyświetlał tylko kraje, dla których są dostępne jakiekolwiek dane.

Poniżej przedstawiono diagram modelu biznesowego w narzędziu Power BI, który opisuje strukturę danych oraz zależności między tabelami i wymiarami w warstwie raportowej.

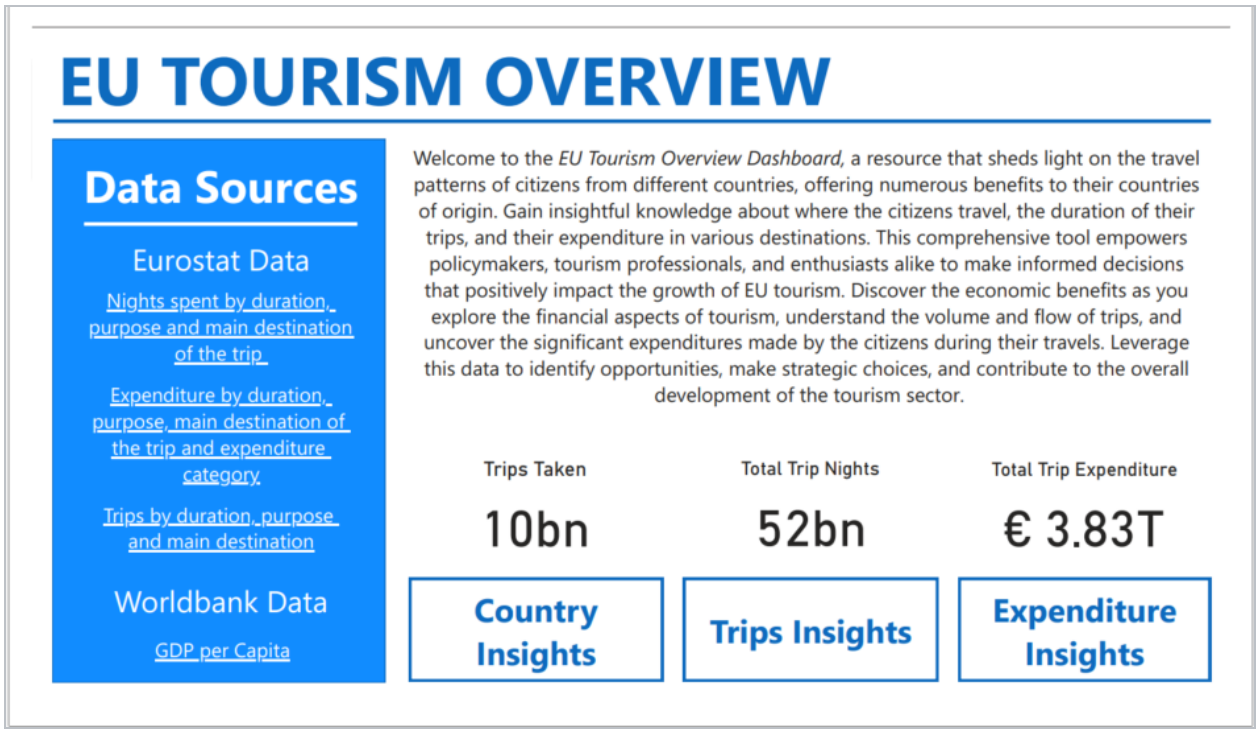


Realizacja przykładowych raportów dla użytkownika

Raport został podzielony na pięć sekcji, z których każda wyjaśnia pewien aspekt turystyki w Unii Europejskiej:

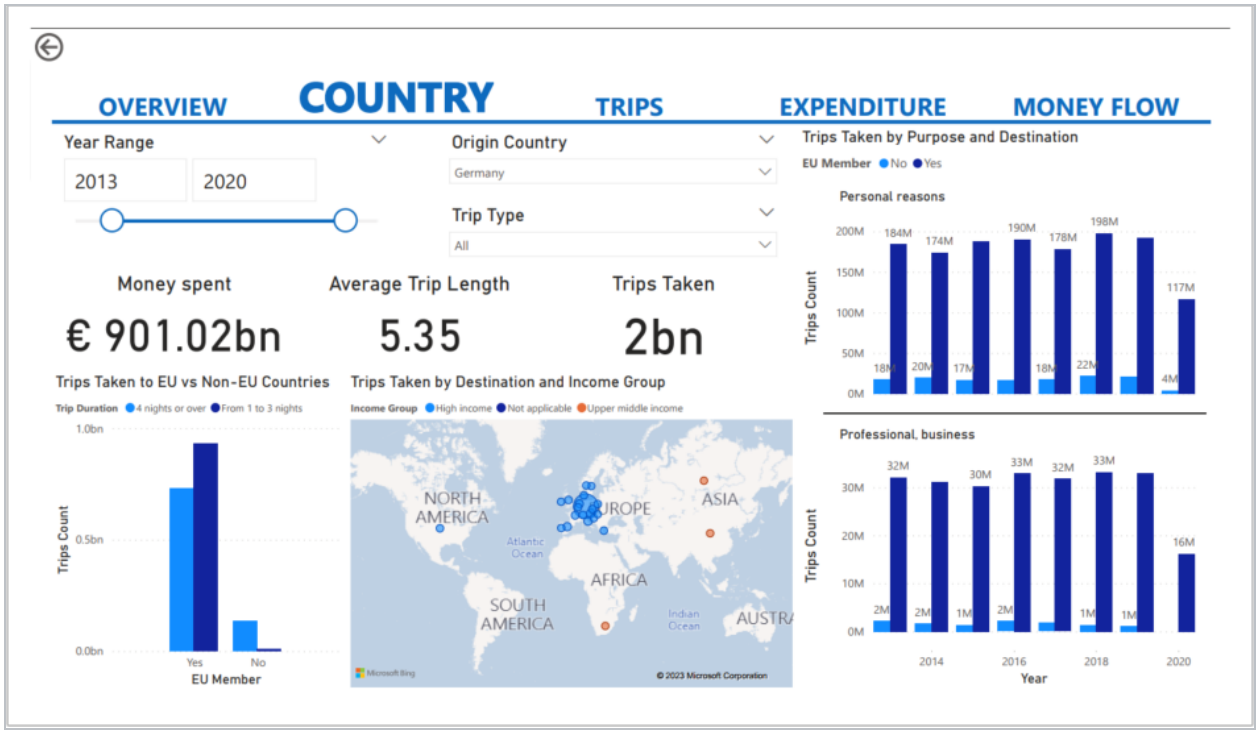
1. Overview:

   Sekcja "Overview" stanowi punkt startowy raportu, mający na celu wprowadzenie użytkownika w tematykę dashboardu. Zawiera również źródła danych użytych przy tworzeniu raportów wraz z odpowiednimi linkami oraz opis, który umożliwia lepsze zrozumienie tematyki dashboardu. Na tej stronie prezentowane są również podstawowe statystyki na najwyższym poziomie, które pozwalają zobaczyć skalę branży. Z sekcji "Overview" istnieje możliwość nawigacji do trzech innych sekcji, z których każda opisuje jeden z aspektów podróżowania i wspomaga lepsze zrozumienie trendów i wzorców zachowań obywateli Unii Europejskiej.



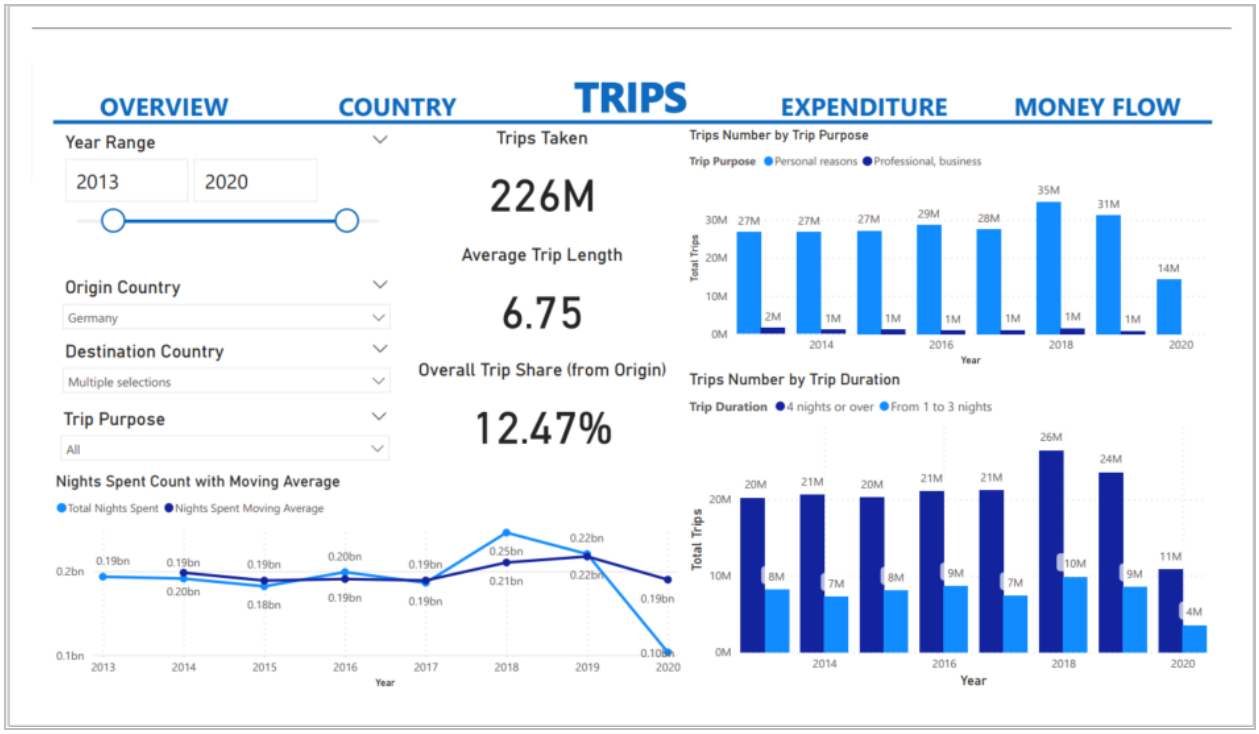
2. Country:

Sekcja "Country" prezentuje szczegółowe informacje dotyczące poszczególnych krajów w Unii Europejskiej związanych z turystyką. Dostępne są dane dotyczące liczby wyjazdów turystycznych, średniej długości wyjazdu oraz wydatków obywateli danego kraju na podróżowanie. Użytkownik ma możliwość wyboru konkretnego kraju, zakresu dat, typu podróży (krajowe, międzynarodowe) i zgłębienia informacji dotyczących turystyki w tym kraju. Wykresy przedstawione na stronie wizualizują podróże z podziałem na cel podróży na przestrzeni lat, uwzględniając kraje będące i niebędące członkami UE, oraz zawierają informacje na temat wydarzeń w poszczególnych latach w podpowiedziach. Ponadto, na stronie znajduje się interaktywna mapa, przedstawiająca wszystkie kraje docelowe zaznaczone w zależności od ich popularności wśród turystów, a także pokolorowane według krajowej grupy dochodowej.



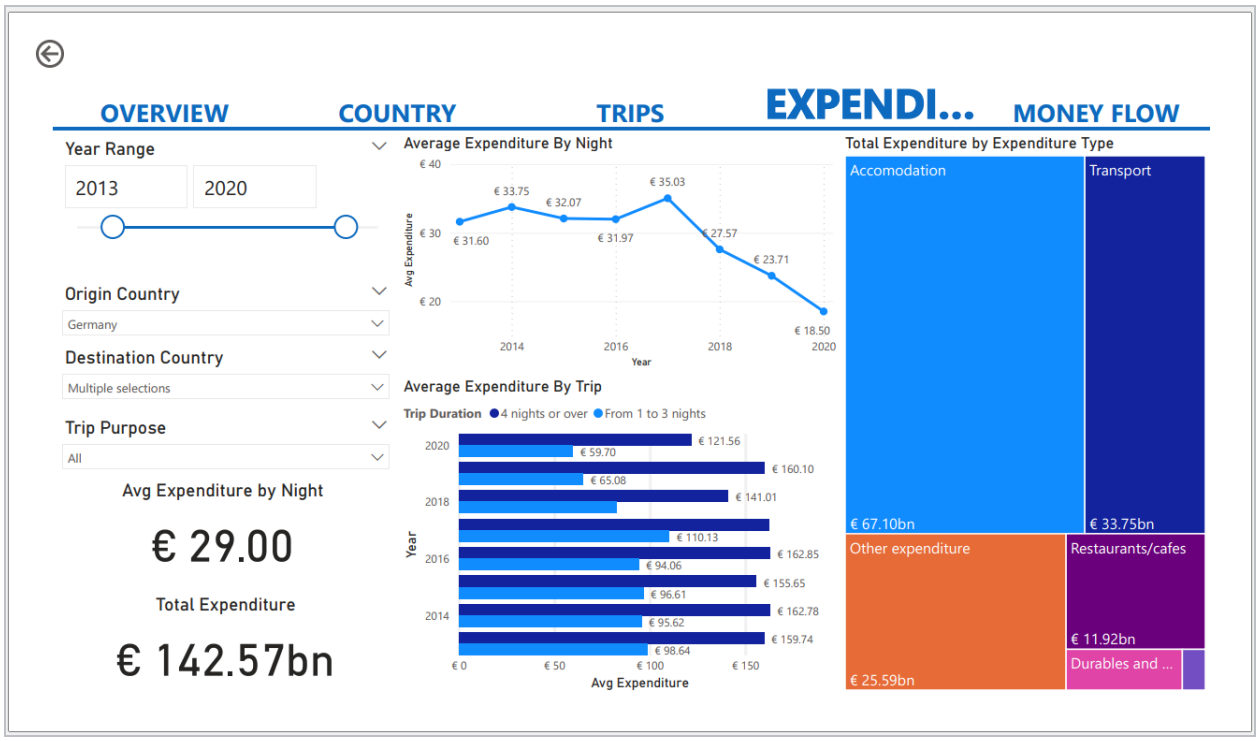
3. Trips:

Sekcja "Trips" koncentruje się na analizie podróży obywateli Unii Europejskiej. Użytkownik ma możliwość zapoznania się z danymi dotyczącymi liczby podróży, takich jak podróże służbowe czy prywatne, zależnie od własnych preferencji i potrzeb. Istnieje możliwość wyboru konkretnego kraju pochodzenia turystów, dowolnej kombinacji krajów/regionów docelowych podróży, zakresu dat oraz celu podróży (prywatne, biznesowe). Dostępne są również dane dotyczące liczby wyjazdów turystycznych, średniej długości wyjazdu oraz procentowego udziału w ogólnej liczbie podróży dla wybranej konfiguracji. Wykresy przedstawiają analizy takie jak ruchoma średnia sumy spędzonych nocy, liczba wycieczek według celu podróży oraz liczba wycieczek podzielonych na kategorie długości wycieczki, prezentowane na przestrzeni lat.



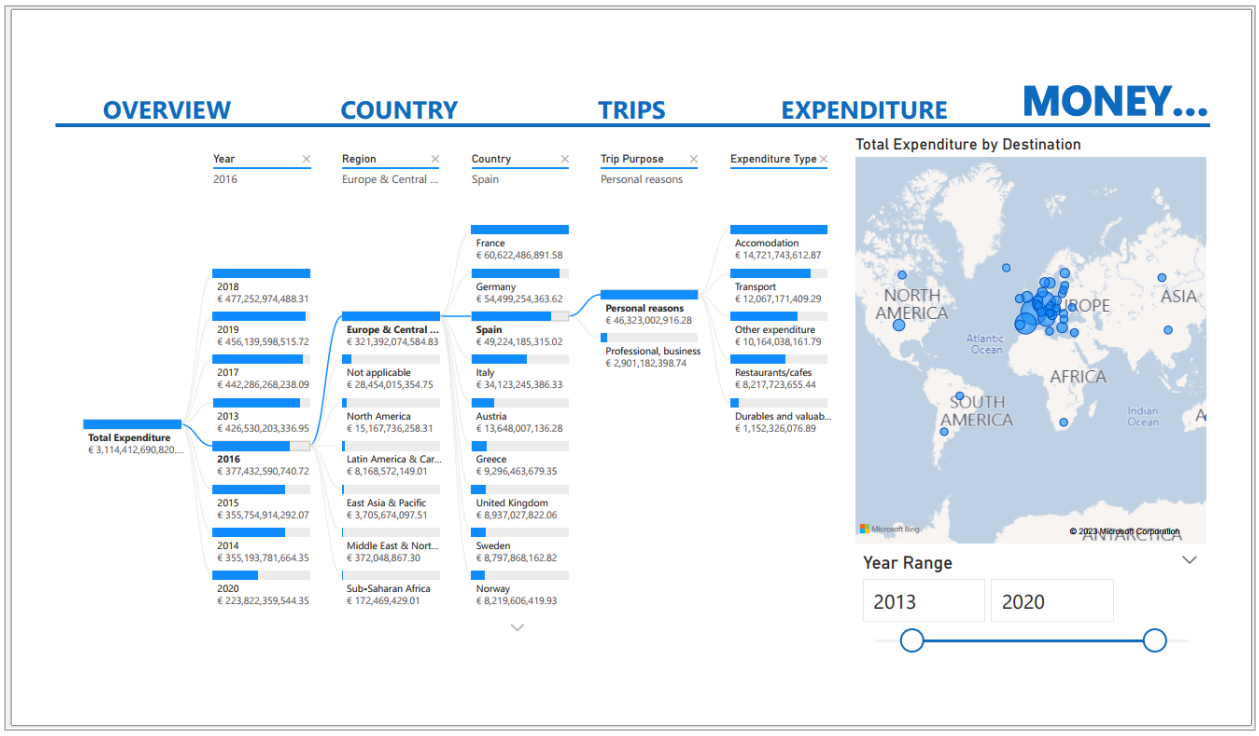
4. Expenditure:

Sekcja "Expenditure" skupia się na analizie wydatków związanych z podróżami w Unii Europejskiej z perspektywy obywateli jednego z krajów UE. Użytkownik ma możliwość wyboru konkretnego kraju pochodzenia turystów, dowolnej kombinacji krajów/regionów docelowych podróży, zakresu dat oraz celu podróży (prywatne, biznesowe). Raport umożliwia zapoznanie się z danymi dotyczącymi średnich wydatków turystów na różne kategorie, takie jak zakwaterowanie, transport, żywność czy atrakcje turystyczne, prezentowane w dostępnych wizualizacjach. Dodatkowo, dostępne są informacje na temat trendów w wydatkach turystycznych, takie jak średni wydatek za noc pobytu oraz całkowite wydatki dla wybranej konfiguracji. Przedstawione są również wizualizacje zawierające informacje dotyczące średnich wydatków za noc pobytu oraz średnich wydatków na całą podróż, z podziałem na kategorie długości wycieczki, prezentowane na przestrzeni czasu.



5. Money flow:

Sekcja "Money flow" prezentuje informacje dotyczące przepływu pieniędzy w sektorze turystycznym Unii Europejskiej. Raport zawiera dane dotyczące wydatków obywateli UE na podróże zagraniczne i wewnątrz kraju, z podziałem na kategorie wydatków. Na stronie znajduje się również interaktywna mapa, umożliwiająca porównanie wydatków w poszczególnych krajach docelowych. Ta strona służy jako dopełnienie do strony “Expenditure” i pomaga w lepszym zrozumieniu analizy przedstawionej na tej stronie.



Każda sekcja raportu dostarcza szczegółowych informacji, wykresów i tabel, które pomagają użytkownikowi zrozumieć i analizować dane dotyczące turystyki w Unii Europejskiej.

Testy pobierania danych

Aby upewnić się, że pobrane dane zostały pomyślnie wczytane do pośredniej bazy danych, przeprowadzony został bardzo prosty test polegający jedynie na podejrzeniu zawartości tabel w pośredniej bazie danych w celu wykrycia oczywistych problemów, takich jak braki kolumn czy umieszczenie wartości w niepoprawnych kolumnach bądź też kompletny brak rekordów w którejś z tabel.

Wynik testu:

```
SELECT * FROM trips
SELECT * FROM nights
SELECT * FROM expenditures
SELECT * FROM countries
```

100 %

Results | Messages

	DATAFLOW	LAST_UPDATE	freq	c_dest	purpose	duration	unit	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
49	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2020	75595	NULL
50	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2021	NULL	u
51	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	E...	2015	288760	NULL
52	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	E...	2016	267847	NULL
53	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	E...	2017	417636	NULL
54	ESTAT:TOUR_DEM_TTW(1.0)	NULL	A	AFR	PER	N1-3	NR	E...	2018	338145	NULL

	DATAFLOW	LAST_UPDATE	freq	c_dest	purpose	duration	unit	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
42	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2013	845338	NULL
43	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2014	796752	NULL
44	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2015	662839	NULL
45	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2016	626156	NULL
46	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2017	998282	NULL
47	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	EA	2018	717701	NULL
48	ESTAT:TOUR_DEM_TNW(1.0)	NULL	A	AFR	PER	N1-3	NR	E...	2019	1032752	NULL

	DATAFLOW	LAST_UPDATE	freq	purpose	duration	c_dest	expend	statinfo	unit	geo	TIME_PERIOD	OBS_VALUE	OBS_FLAG
5	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2012	229.80	NULL
6	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2013	228.20	NULL
7	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2014	188.01	NULL
8	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2015	210.64	NULL
9	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2016	170.47	NULL
10	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2017	115.91	NULL
11	ESTAT:TOUR_DEM_EXTOTW(1.0)	NULL	A	TOTAL	N_GE4	NO	ACCOM	AVG_TRP	EUR	LT	2018	160.75	NULL

	Country_Code	Country_Name	Income_Group	Region	Country_Long_Name
1	AW	Aruba	High income	Latin America & Caribbean	Aruba
2	ZH	Africa Eastern and Southern	NULL	NULL	Africa Eastern and Southern
3	AF	Afghanistan	Low income	South Asia	Islamic State of Afghanistan
4	A9	Africa	NULL	NULL	Africa
5	ZI	Africa Western and Central	NULL	NULL	Africa Western and Central
6	AO	Angola	Lower middl...	Sub-Saharan Africa	People's Republic of Ango...

Query executed successfully. | DESKTOP-2ISVH1V (15.0 RTM) | DESKTOP-2ISVH1V\macie ... | kapidane_raw | 00:00:19 | 190,658 rows

Wynik testu wskazuje na brak oczywistych niezgodności po wczytaniu danych.

Uznaliśmy, że bardziej dogłębne testowanie tych danych nie ma z naszej strony sensu, gdyż wówczas tak naprawdę testami wyszlibyśmy poza nasz system - ewentualne problemy sugerowałyby problemy po stronie API dostawcy danych, których i tak nie bylibyśmy w stanie rozwiązać.

Testy jakości danych

Aby zapewnić wiarygodność i dokładność danych wykorzystywanych w tym projekcie, został wdrożony kompleksowy zestaw testów jakości danych. Oceniają one różne aspekty jakości danych, w tym kompletność, dokładność, spójność, ważność, integralność, aktualność, powielanie i niepowtarzalność. Testy jakości danych obejmują:

- Test kompletności, aby sprawdzić, czy wszystkie wymagane pola są wypełnione dla każdego rekordu, zidentyfikować brakujące lub puste wartości w krytycznych polach danych, sprawdzić kompletność danych we wszystkich wymiarach i miarach.
- Test ważności w celu sprawdzenia zgodności danych z predefiniowanymi typami danych, formatami i zakresami wartości.
- Test powielania w celu zidentyfikowania zduplikowanych rekordów lub wpisów danych, które mogą prowadzić do zawyżonych lub nieprawidłowych wyników.

Testy zostały zgrupowane według wymiaru lub tabeli faktów do której się odnoszą. Dla każdej tabeli stworzonych zostało od dwóch do pięciu testów (w zależności od złożoności danego wymaru / tabeli faktów). Cały zestaw testów jakości danych załączony jest wraz z kodem źródłowym w folderze *tests/ssis*.

Testy wymiarów Dim_TripDetails oraz Dim_ExpenditureDetails

Na tych dwóch wymiarach przeprowadzane testy były do siebie analogiczne, z tego powodu zostaną opisane one jednocześnie. Jako, że te wymiary są stosunkowo niewielkie (mają po kilka rekordów), stworzone zostały dla nich tylko dwa testy.

1. Test poprawności ładowania danych.

   Test ten sprawdza czy dane załadowane do hurtowni są kompletne. W tym celu poszukiwane są wszystkie unikalne wartości z odpowiednich kolumn w danych źródłowych i porównywane są one z zawartością wymiarów. Dla Dim_TripDetails są to kolumny *purpose* i *duration* z tabel źródłowych *trips*, *nights* oraz *expendiures*, zaś dla Dim_ExpenditureDetails jest to kolumna *expend* z tabeli *expenditures*. Oczekiwane jest, że dane z tabel źródłowych będą pokrywać się z danymi z hurtowni. Dodatkowo, doklejane są mapowania z tabel wymiarów aby móc zweryfikować ich poprawność.

   Wynik dla Dim_TripDetails:

	source	target	TripPurpose	TripDuration
1	PERN1-3	PERN1-3	Personal reasons	From 1 to 3 nights
2	PERN_GE4	PERN_GE4	Personal reasons	4 nights or over
3	PROFN1-3	PROFN1-3	Professional, business	From 1 to 3 nights
4	PROFN_GE4	PROFN_GE4	Professional, business	4 nights or over

Wynik dla Dim_ExpenditureDetails:

	source	target	ExpenditureType
1	ACCOM	ACCOM	Accomodation
2	DUR	DUR	Durables and valuable goods
3	PACK_ARR	PACK_ARR	Package arrangement
4	REST	REST	Restaurants/cafes
5	TRA	TRA	Transport
6	TRP_OTH	TRP_OTH	Other expenditure

Wyniki potwierdzają zgodność danych z tabelami źródłowymi.

2. Test poprawności dodawania nowych danych.

   Test ten ma na celu sprawdzić, czy w razie dodania w tabeli źródłowej nowej kategorii wycieczki bądź też kategorii wydatku odpowiednie wymiary będą się poprawnie aktualizować. Aby zasymulować takie wydarzeniue, z tabel wymiarów usuwany jest jeden (lub więcej) rekord, a następnie uruchamiany jest proces wypełniania wymiarów w SSIS. Oczekiwane jest, że po jego zakończeniu stan tabel wymiarów będzie taki sam jak przed usunięciem rekordu/rekordów.

   Wynik dla Dim_TripDetails:

	TripDetailsKey	TripPurpose	TripDuration
1	PERN_GE4	Personal reasons	4 nights or over
2	PERN1-3	Personal reasons	From 1 to 3 nights
3	PROFN_GE4	Professional, business	4 nights or over
4	PROFN1-3	Professional, business	From 1 to 3 nights

   Wynik dla Dim_ExpenditureDetails:

	ExpenditureDetailsKey	ExpenditureType
1	ACCOM	Accomodation
2	DUR	Durables and valuable goods
3	PACK_ARR	Package arrangement
4	REST	Restaurants/cafes
5	TRA	Transport
6	TRP_OTH	Other expenditure

   Wynikowe stany wymiarów potwierdzają, że brakujące rekordy zostały poprawnie uzupełnione.

Testy wymiaru Dim_Geography

1. Test poprawności ładowania danych.

   Jest to prosty test mający na celu wykryć oczywiste niezgodności po wczytaniu danych. Polega on jedynie na podejrzeniu zawartości tabeli wymiarowej i

zweryfikowaniu, czy dane w niej na pierwszy rzut oka wyglądają zgodnie z oczekiwaniami. Przykładem niepoprawnego wczytania danych mogłaby być np. sytuacja, gdy wartości jednej z kolumn byłyby zamienione z drugą.

Wynik testu:

	CountryKey	CountryCode	CountryName	Region	IncomeGroup	IsMemberStateEU	ValidFromDate	ValidToDate	IsCurrent
1	1	AFR_OTH	Other African countries	Not applicable	Not applicable	No	2023-06-11	NULL	Yes
2	2	AME_C_S_OTH	Other Central or South American countries	Latin America & Caribbean	Not applicable	No	2023-06-11	NULL	Yes
3	3	AME_N_OTH	Other North American countries	North America	Not applicable	No	2023-06-11	NULL	Yes
4	4	ASI_OTH	Other Asian countries	Not applicable	Not applicable	No	2023-06-11	NULL	Yes
5	5	EUR_OTH	Other European countries	Europe & Central Asia	Not applicable	No	2023-06-11	NULL	Yes
6	6	OCE_OTH	Other Oceanian countries	Not applicable	Not applicable	No	2023-06-11	NULL	Yes
7	7	AW	Aruba	Latin America & Caribbean	High income	No	2023-06-11	NULL	Yes
8	8	AF	Afghanistan	South Asia	Low income	No	2023-06-11	NULL	Yes
9	9	AO	Angola	Sub-Saharan Africa	Lower middle income	No	2023-06-11	NULL	Yes
10	10	AL	Albania	Europe & Central Asia	Upper middle income	No	2023-06-11	NULL	Yes

Wynik potwierdza brak oczywistych niezgodności po wczytaniu danych do wymiaru.

2. Test kompletności danych.

   Ma on za zadanie zweryfikowanie, czy podczas wczytywania danych ze źródła do hurtowni nie doszło do utraty żadnych rekordów. W tym celu zliczana jest liczba rekordów w tabelach źródłowych oraz liczba aktualnych (SCD2) rekordów w tabeli wymiaru. Osobno zliczane są kraje i grupy krajów (czyli kraje z kodem kraju zakończonym na “OTH”), ponieważ informacje o nich ładowane są z osobnych źródeł. Porównywane są też zagregowane wartości z niektórych kolumn. Oczekiwane jest, że liczby rekordów w źródłach i w wymiarze oraz zagregowane wartości wybranych kolumn będą sobie równe.

   Wynik testu:

	(No column name)	(No column name)	CountriesCount	Max(IncomeGroup)	Min(Region)
1	Countries	Source	217	Upper middle income	East Asia & Pacific
2	Countries	Target	217	Upper middle income	East Asia & Pacific
3	Country groups	Source	6	Not applicable	Europe & Central Asia
4	Country groups	Target	6	Not applicable	Europe & Central Asia

   Wynik potwierdza kompletność danych z hurtowni.

3. Test poprawności dodawania nowych danych.

   Test ten ma na celu weryfikację czy w razie dodania w danych źródłowych nowego kraju do hurtowni poprawnie zostanie dodany nowy rekord. Test ten jest analogiczny jak w przypadku Dim_TripDetails i Dim_Expenditures, tzn. usuwany jest jeden rekord i sprawdzane jest, czy po powtórnym załadowaniu wymiarów zostanie on z powrotem dodany. W ramach testu usunięty został rekord odpowiadający Irlandii. Oczekiwane jest, że po przeprowadzeniu testu będzie on ponownie znajdował się w wymiarze.

   Wynik testu:

	CountryKey	CountryCode	CountryName	Region	IncomeGroup	IsMemberStateEU	ValidFromDate	ValidToDate	IsCurrent
1	226	IE	Ireland	Europe & Central Asia	High income	Yes	2023-06-11	NULL	Yes

Usunięty uprzednio rekord znajduje się w wymiarze, co potwierdza poprawność ładowania nowych danych do wymiaru.

4. Test działania SCD2

   Test ten ma na celu sprawdzenie, czy zaimplementowany mechanizm SCD typu 2 działa zgodnie z oczekiwaniami. Zweryfikować to można dość prosto poprzez podmianę wartości np. z kolumny IncomeGroup dla dowolnego kraju i ponowne załadowanie wymiarów i skrypt do takiego scenariusza zawarty jest w scenariuszach testowych załączonych do kodu źródłowego. Tutaj jednak działanie SCD2 zaprezentowane zostanie na ciekawszym i bardziej złożonym przykładzie - na uwzględnianiu zmian członkowskich w Unii Europejskiej na przykładzie UK.

   Scenariusz tego testu wraz ze skryptami opisany jest dokładnie w pliku *scripts/handling_UK.sql*. W pierwszej kolejności należy upewnić się, że w tabeli pomocniczej Lookup_EU w hurtowni znajduje się rekord odpowiadający UK. Następnym krokiem jest załadowanie wymiarów oraz faktów do roku 2020. Dalej, ze wcześniej wspomnianej tabeli Lookup_EU należy usunąć rekord odpowiadający UK. Potem należy ponownie załadować wymiary, a także fakty, tym razem jednak od roku 2021. Oczekiwane jest że w Dim_Geography będą znajdowały się 2 rekordy odpowiadające UK oraz że fakty z różnych lat będą przypisane do odpowiednich rekordów z wymiaru geograficznego.

   Wyniki testu:

	CountryKey	CountryCode	CountryName	Region	IncomeGroup	IsMemberStateEU	ValidFromDate	ValidToDate	IsCurrent
1	12	UK	United Kingdom	Europe & Central Asia	High income	Yes	2023-06-11	2023-06-11	No
2	224	UK	United Kingdom	Europe & Central Asia	High income	No	2023-06-11	NULL	Yes

	CountryName	YearKey	IsMemberStateEU	ValidFromDate	ValidToDate	IsCurrent
1	United Kingdom	2012	Yes	2023-06-11	2023-06-11	No
2	United Kingdom	2013	Yes	2023-06-11	2023-06-11	No
3	United Kingdom	2014	Yes	2023-06-11	2023-06-11	No
4	United Kingdom	2015	Yes	2023-06-11	2023-06-11	No
5	United Kingdom	2016	Yes	2023-06-11	2023-06-11	No
6	United Kingdom	2017	Yes	2023-06-11	2023-06-11	No
7	United Kingdom	2018	Yes	2023-06-11	2023-06-11	No
8	United Kingdom	2019	Yes	2023-06-11	2023-06-11	No
9	United Kingdom	2020	Yes	2023-06-11	2023-06-11	No
10	United Kingdom	2021	No	2023-06-11	NULL	Yes

   Wyniki potwierdzają poprawność działania mechanizmu SCD2.

Testy faktów Fact_TripsNights oraz Fact_Expenditures

Na obu tabelach faktowych przeprowadzane były analogiczne testy, stąd opisane one zostaną w jednym miejsu, a wyniki będą prezentowane równolegle.

1. Test poprawności ładowania danych.

   Test ten działa dokładnie tak samo jak w przypadku wymiaru geograficznego, tzn. polega jedynie na wglądzie w dane w celu sprawdzenia czy nie ma w nich oczywistych niezgodności. Oczekiwane jest, że dane na pierwszy rzut oka wyglądać będą zgodnie z oczekiwaniami.

   Wynik testu dla Fact_TripsNights:

	FactKey	YearKey	TripDetailsKey	OriginCountryKey	DestinationCountryKey	TripsCount	NightsSpentCount
1	2012ATAFR_OTHPERN_GE4	2012	PERN_GE4	1	29	172937	1702965
2	2012ATAME_C_S_OTHPERN_GE4	2012	PERN_GE4	1	30	70529	1348850
3	2012ATASI_OTHPERN_GE4	2012	PERN_GE4	1	32	394107	4854569
4	2012ATASI_OTHPROFN_GE4	2012	PROFN_GE4	1	32	55030	489832
5	2012ATATPERN_GE4	2012	PERN_GE4	1	1	3238081	21630813
6	2012ATATPERN1-3	2012	PERN1-3	1	1	5777112	11614059
7	2012ATATPROFN_GE4	2012	PROFN_GE4	1	1	229407	1751660
8	2012ATATPROFN1-3	2012	PROFN1-3	1	1	1717256	2687934
9	2012ATCHPERN_GE4	2012	PERN_GE4	1	65	120590	952870

   Wynik testu dla Fact_Expenditures:

	FactKey	YearKey	TripDetailsKey	OriginCountryKey	DestinationCountryKey	ExpenditureDetailsKey	AvgExpenditureByNight	AvgExpenditureByTrip
1	2012ATAFR_OTHPERN_GE4ACCOM	2012	PERN_GE4	1	29	ACCOM	77.66	764.73
2	2012ATAFR_OTHPERN_GE4DUR	2012	PERN_GE4	1	29	DUR	1.56	15.31
3	2012ATAFR_OTHPERN_GE4TRA	2012	PERN_GE4	1	29	TRA	43.33	426.69
4	2012ATAFR_OTHPERN_GE4TRP_OTH	2012	PERN_GE4	1	29	TRP_OTH	12.99	127.87
5	2012ATAME_C_S_OTHPERN_GE4ACCOM	2012	PERN_GE4	1	30	ACCOM	60.05	1148.35
6	2012ATAME_C_S_OTHPERN_GE4DUR	2012	PERN_GE4	1	30	DUR	1.52	29.06
7	2012ATAME_C_S_OTHPERN_GE4TRA	2012	PERN_GE4	1	30	TRA	46.79	894.93
8	2012ATAME_C_S_OTHPERN_GE4TRP_OTH	2012	PERN_GE4	1	30	TRP_OTH	17.88	342.02

   Wyniki potwierdzają brak oczywistych niezgodności załadowanych danych.

2. Test poprawności tworzenia TripDetailsKey

   Celem testu jest zweryfikowanie czy transformacja odpowiedzialna za stworzenie klucza obcego do wymiaru Dim_TripDetails działa poprawnie. Aby to sprawdzić, konkatenowane są wszystkie parami unikalne wartości kolumn *purpose* i *duration* z dowolnej z tabel źródłowych. Dalej wybierane są wszystkie unikalne wartości z kolumny TripDetailsKey w wymiarze. Na koniec dokonywany jest FULL OUTER JOIN obu tych zestawów wartości. Jeśli klucz stworzony został niepoprawnie, to w tabeli zwracanej przez test będą znajdować się wartości NULL.

   Wynik testu (dla obu tabel był identyczny):

	source	target
1	PERN_GE4	PERN_GE4
2	PROFN_GE4	PROFN_GE4
3	PERN1-3	PERN1-3
4	PROFN1-3	PROFN1-3

   Brak wartości NULL w powyższej tabeli potwierdza poprawność tworzenia tego klucza.

3. Test zgodności referencji do wymiarów

   Test ten ma za zadanie zbadać, czy każda z wartości kluczów obcych w tabalach faktów faktycznie odnosi się do istniejących rekordów w tabelach wymiarów. Aby to zbadać, dokonywany jest INNER JOIN tabel faktów ze wszystkimi wymiarami, a następnie zliczana jest liczba wierszy w otrzymanej tabeli. Oczekiwane jest, że liczba ta będzie równa liczbie wierszy w tabeli faktów.

   Wynik dla Fact_TripsNights:

	(No column name)	Count
1	Standalone fact table	9230
2	Fact table joined with dimensions	9230

   Wynik dla Fact_Expenditures:

	(No column name)	Count
1	Standalone fact table	45373
2	Fact table joined with dimensions	45373

   Wyniki potwierdzają zgodność wszystkich referencji do wymiarów.

4. Test unikalności faktów

   Zadaniem tego testu jest upewnenie się, że w tabelach faktowych nie ma zduplikowanych faktów, tzn. dwóch lub więcej faktów mających identyczne klucze obce do wymiarów (same wartości faktów mogą być równe lub nie). Najpierw uruchamiany jest proces ładowania faktów przynajmniej dwukrotnie, z nakładającymi się na siebie zakresami lat. Następnie zliczane są wszystkie rekordy w tabeli faktów z unikalnymi kombinacjami wartości kluczy obcych. Oczekiwane jest, że ta liczba będzie równa liczbie rekordów w pełnej tabeli faktów.

   Wynik dla Fact_TripsNights:

	(No column name)	Count
1	All facts	9230
2	Unique facts	9230

   Wynik dla Fact_Expenditures:

	(No column name)	Count
1	All facts	45373
2	Unique facts	45373

   Wyniki testów potwierdzają unikalność faktów w obu tabelach.

5. Test ważności danych

   Test ten ma za zadanie porównać liczbę rekordów oraz zakresy wartości miarek w tabelach faktów w hurtowni z tabelami źródłowymi. W tym celu porównywane są te statystyki i oczekiwane jest, że będą one równe pomiędzy pomiędzy źródłem i tabelami docelowymi.

   Wynik dla Fact_TripsNights:

	(No column name)	Count	MaxTripsCount	MinTripsCount	MinNightsSpentCount	MinNightsSpentCount
1	source	9230	101090835	1066	790872686	2567
2	target	9230	101090835	1066	790872686	2567

   Wynik dla Fact_Expenditures:

	(No column name)	Count	MaxAvgExpendPerTrip	MinAvgExpendPerTrip	MaxAvgExpendPerNight	MinAvgExpendPerNight
1	source	45373	3388.22	0.00	1620.88	0.00
2	target	45373	3388.22	0.00	1620.88	0.00

   Wyniki potwierdzają ważność danych w tabelach faktów.

Testy wizualizacji

Aby zapewnić poprawność danych znajdujących się na raportach, stworzony został zestaw testów weryfikujący poprawność wizualizacji w narzędziu BI. Testy te porównują wartości przedstawione na wizualizacji z faktycznymi wartościami w samej hurtowni danych, otrzymywanymi poprzez odpowiednie zapytania SQL. Dla każdej wizualizacji oraz KPI został stworzony oddzielny test. Cały zestaw testów wizualizacji znajduje się w folderze *tests/reports*.

Strona "Overview"

1. Test KPI

   Celem testu było sprawdzenie zgodności wyświetlanych wartości KPI z wartościami uzyskanymi poprzez zapytanie SQL bezpośrednio na danych z hurtowni. Oczekiwane jest, że te wartości będą równe (tak naprawdę w przybliżeniu równe, gdyż wartości na raporcie dla czytelności i estetyki są zaokrąglane)

   Wartości w raporcie:

Trips Taken	Total Trip Nights	Total Trip Expenditure
10bn	52bn	€ 3.83T

Wartości w hurtowni:

	Trips Taken	Total Trip Nights	Total Trip Expenditure
1	10420191669	52445661352	3831011101720.04

Wyniki testu potwierdzają poprawność danych w raporcie.

Strona "Country"

W trakcie testów na tej stronie zastosowane były następujące filtry:

- Year Range: 2013 - 2020,
- Origin Country: Germany,
- Trip Type: All.

1. Test KPI

   Wartości w raporcie:

Money spent	Average Trip Length	Trips Taken
€ 901.02bn	5.35	2bn

   Wartości w hurtowni:

	(No column name)	(No column name)
1	Money Spent	901015384360.93
2	Average Trip Length	5.34930717353713
3	Trips Taken	1809225206

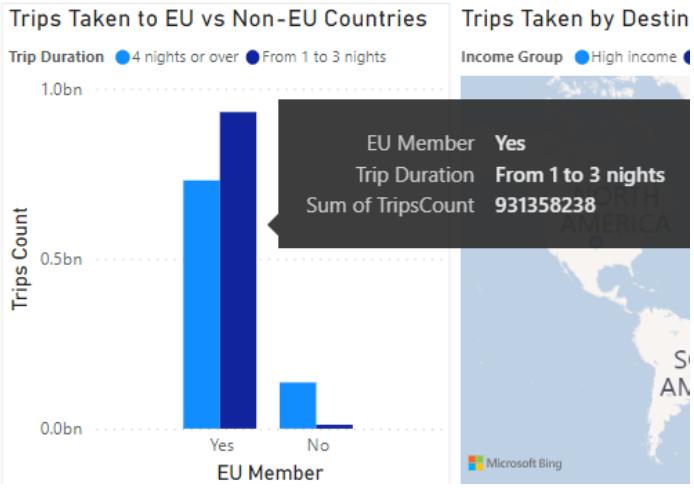
   Wartości wyświetlane na raporcie są w rzeczy samej zaokrągleniami[1] wartości uzyskanych poprzez odpowiednie zapytanie w hurtowni, co dowodzi poprawność danych w raporcie.

2. Test wizualizacji: wykres słupkowy "Trips taken to EU vs Non-EU countries"

   Test miał za zadanie zweryfikować czy wartości wyświetlane na wykresie są zgodne z rzeczywistym stanem hurtowni. W tym celu wykonane zostało zapytanie SQL bezpośrednio na hurtowni. Oczekiwane jest, że te dane będą pokrywać się z danymi widocznymi na wykresie.

[1] O ile zastanawiać może jedynie wyświetlana wartość 2bn w porównaniu z rzeczywistą wartością równą 1.8bn w hurtowni, to gdyby usunąć zaokrąglanie z wizualizacji to wówczas wartości faktycznie się pokrywają

Wartości na wykresie:



Wartości uzyskane z hurtowni:

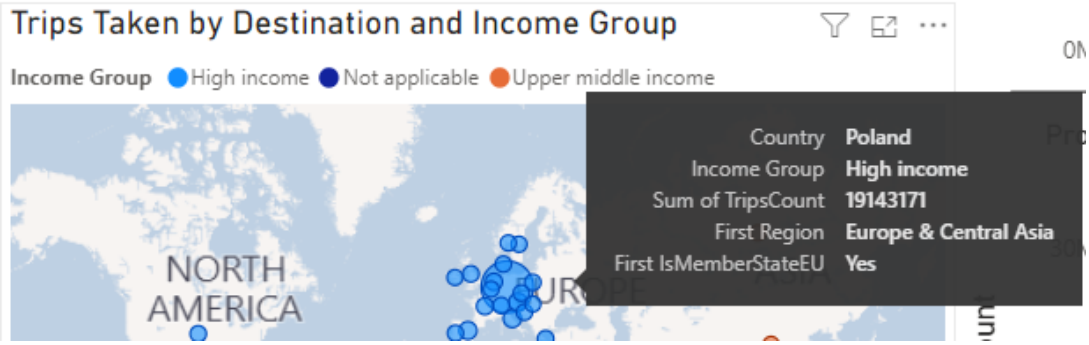
	IsMemberStateEU	TripDuration	(No column name)
1	Yes	4 nights or over	730545365
2	Yes	From 1 to 3 nights	931358238
3	No	4 nights or over	135850456
4	No	From 1 to 3 nights	11471147

Wartości uzyskane na oba sposoby pokrywają się co dowodzi poprawności wizualizacji.

3. Test wizualizacji: mapa “Trips Taken by Destination and Income Group”

Celem testu było zweryfikowanie czy wartości wyświetlane na mapie pokrywają się ze stanem hurtowni. W tym celu wykonane zostało na hurtowni zapytanie SQL zwracające takie dane, jakie oczekuje się, że będą znajdować się na mapie.

Przykładowa wartość na mapie:



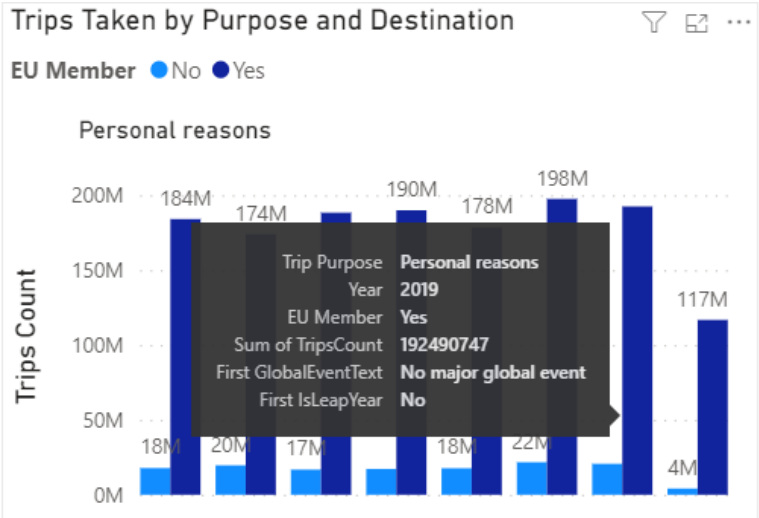
Wartości uzyskane z hurtowni:

19	Yes	High income	Europe & Central Asia	Poland	19143171

Porównywane wartości są równe co potwierdza poprawność wizualizacji.

4. Test wizualizacji: wykres słupkowy "Trips Taken By Purpose and Destination - Personal Reasons"

Wartości na wykresie:



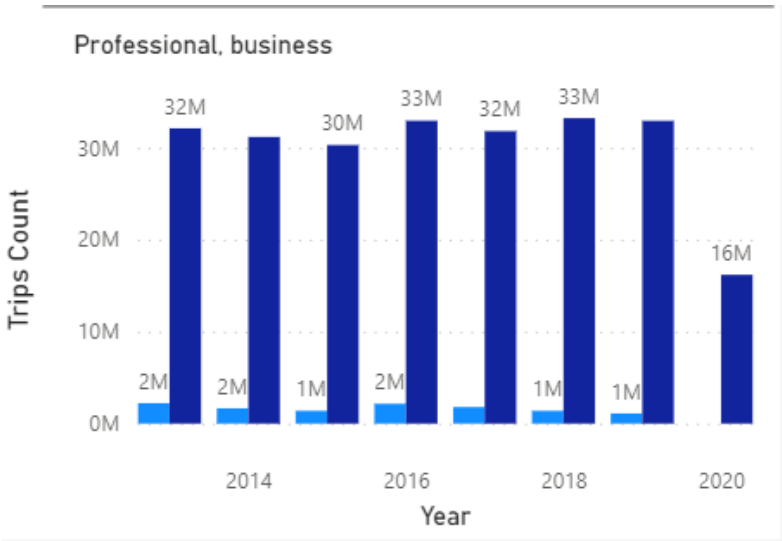
Wartości w hurtowni:

	TripPurpose	YearKey	IsMemberStateEU	GlobalEventText	(No column name)
1	Personal reasons	2013	No	No major global event	17811172
2	Personal reasons	2013	Yes	No major global event	184071606
3	Personal reasons	2014	No	No major global event	19618180
4	Personal reasons	2014	Yes	No major global event	173827890
5	Personal reasons	2015	No	No major global event	16794887
6	Personal reasons	2015	Yes	No major global event	188327651
7	Personal reasons	2016	No	No major global event	17226931
8	Personal reasons	2016	Yes	No major global event	189949962
9	Personal reasons	2017	No	No major global event	17680117
10	Personal reasons	2017	Yes	No major global event	178295016
11	Personal reasons	2018	No	No major global event	21684981
12	Personal reasons	2018	Yes	No major global event	197529756
13	Personal reasons	2019	No	No major global event	20646833
14	Personal reasons	2019	Yes	No major global event	192490747
15	Personal reasons	2020	No	COVID-19 Pandemic	4120960
16	Personal reasons	2020	Yes	COVID-19 Pandemic	116745085

Wartości na wykresie i w hurtowni pokrywają się, co oznacza, że wizualizacja działa poprawnie.

5. Test wizualizacji: wykres słupkowy “Trips Taken By Purpose and Destination - Professional, business”

   Wartości na wykresie:



   Wartości uzyskane bezpośrednio z hurtowni:

17	Professional, business	2013	No	No major global event	2224234
18	Professional, business	2013	Yes	No major global event	32115291
19	Professional, business	2014	No	No major global event	1667521
20	Professional, business	2014	Yes	No major global event	31176888
21	Professional, business	2015	No	No major global event	1391108
22	Professional, business	2015	Yes	No major global event	30292506
23	Professional, business	2016	No	No major global event	2161839
24	Professional, business	2016	Yes	No major global event	32940111
25	Professional, business	2017	No	No major global event	1798341
26	Professional, business	2017	Yes	No major global event	31807910
27	Professional, business	2018	No	No major global event	1391014
28	Professional, business	2018	Yes	No major global event	33222216
29	Professional, business	2019	No	No major global event	1103485
30	Professional, business	2019	Yes	No major global event	32929077
31	Professional, business	2020	Yes	COVID-19 Pandemic	16181891

   Ponownie oba zestawy wartości są równe, co dowodzi zgodności danych wyświetlanych na wizualizacji.

Strona “Trips”

W trakcie testów na tej stronie zastosowane były następujące filtry:

- Year Range: 2013 - 2020;
- Origin Country: Germany;
- Destination Country: Austria, Belgium, Croatia, Czechia, Denmark, Italy, Switzerland;
- Trip Type: All.

1. Test KPI

   Wartości wyświetlane na wizualizacji:

   Trips Taken

   226M

   Average Trip Length

   6.75

   Overall Trip Share (from Origin)

   12.47%

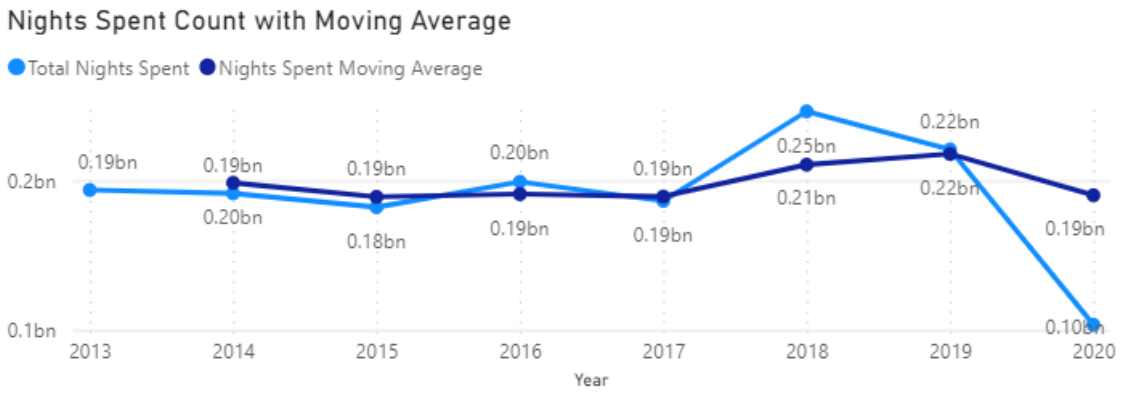
   Wartości uzyskane poprzez zapytania na hurtowni:

	(No column name)	(No column name)
1	Trips taken	225598710
2	Average trip length	6.74706843846758
3	Overall trip share in % (from origin)	12.4693547962873

   Wartości KPI na raporcie pokrywają się z tymi z hurtowni co dowodzi ich poprawności.

2. Test wizualizacji: wykres liniowy “Nights Spent Count with Moving Average”

   Wizualizacja w raporcie:



Wartości uzyskane z hurtowni:

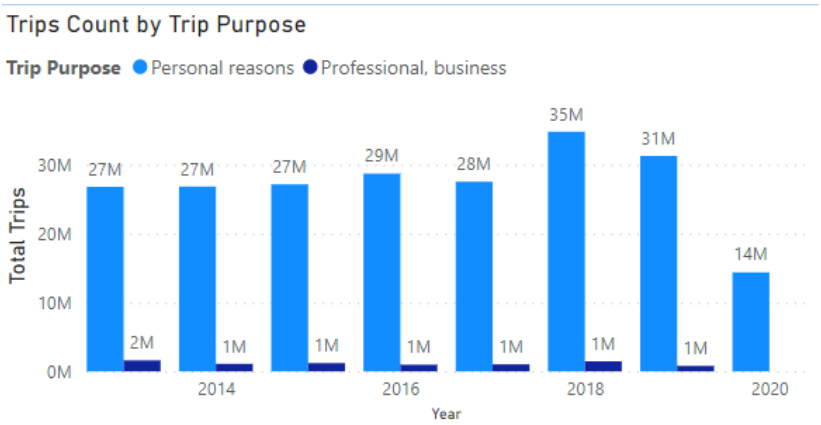
	YearKey	NightsSpentCount
1	2013	193509600
2	2014	191287164
3	2015	182032179
4	2016	199006401
5	2017	186276388
6	2018	246086076
7	2019	220586988
8	2020	103345140

	Year	NightsCountMovingAvg
1	2014	198312323
2	2015	188942981
3	2016	190775248
4	2017	189104989
5	2018	210456288
6	2019	217649817
7	2020	190006068

Wartości z wykresu zgadzają się z tymi uzyskanymi poprzez zapytania na hurtowni, co dowodzi poprawności wizualizacji.

3. Test wizualizacji: wykres słupkowy “Trip Count by Trip Purpose”

Wartości na wykresie:



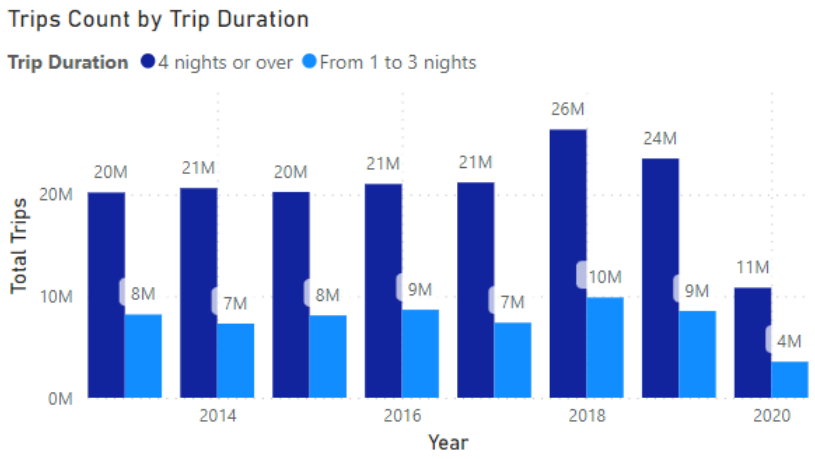
Wartości z hurtowni:

	YearKey	TripPurpose	TripsCount
1	2013	Personal reasons	26792717
2	2013	Professional, business	1592334
3	2014	Personal reasons	26828806
4	2014	Professional, business	1090307
5	2015	Personal reasons	27155421
6	2015	Professional, business	1196033
7	2016	Personal reasons	28717049
8	2016	Professional, business	965665
9	2017	Personal reasons	27544364
10	2017	Professional, business	1021425
11	2018	Personal reasons	34792320
12	2018	Professional, business	1459714
13	2019	Personal reasons	31283342
14	2019	Professional, business	788031
15	2020	Personal reasons	14371182

Wartości umieszczone na wykresie faktycznie zgadzają się z tymi uzyskanymi z hurtowni (te wypisane nad słupkami są zaokrągleniami wartości dokładnych) co potwierdza poprawność wizualizacji.

4. Test wizualizacji: wykres słupkowy “Trip Count by Trip Duration”

   Wizualizacja:



   Wartości pozyskane z hurtowni:

	YearKey	TripDuration	TripsCount
1	2013	4 nights or over	20190264
2	2013	From 1 to 3 nights	8194787
3	2014	4 nights or over	20622846
4	2014	From 1 to 3 nights	7296267
5	2015	4 nights or over	20245056
6	2015	From 1 to 3 nights	8106398
7	2016	4 nights or over	21023762
8	2016	From 1 to 3 nights	8658952
9	2017	4 nights or over	21179700
10	2017	From 1 to 3 nights	7386089
11	2018	4 nights or over	26382807
12	2018	From 1 to 3 nights	9869227
13	2019	4 nights or over	23531096
14	2019	From 1 to 3 nights	8540277
15	2020	4 nights or over	10824770
16	2020	From 1 to 3 nights	3546412

Strona “Expenditures”

W trakcie testów na tej stronie zastosowane były następujące filtry:

- Year Range: 2013 - 2020;
- Origin Country: Germany;
- Destination Country: Austria, Belgium, Croatia, Czechia, Denmark, Italy, Switzerland;
- Trip Type: All.

1. Test KPI

   Wartości na raporcie:

   Avg Expenditure by Night

   € 29.00

   Total Expenditure

   € 142.57bn

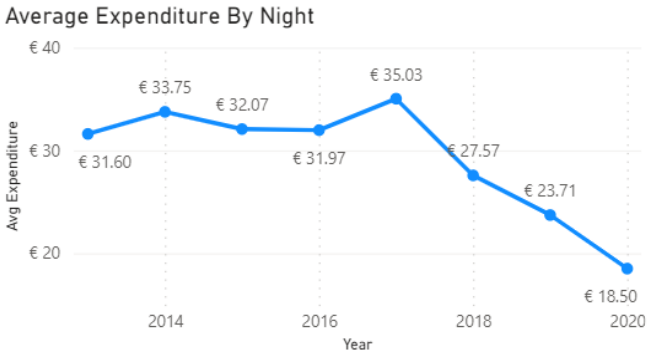
   Wartości w hurtowni:

	(No column name)	(No column name)
1	Avg Expenditure by Night	29.0044
2	Total Expenditure	142568061444.98

   Wyniki potwierdzają poprawność danych w raporcie.

2. Test wizualizacji: wykres liniowy “Average Expenditure By Night”

   Wartości na raporcie:



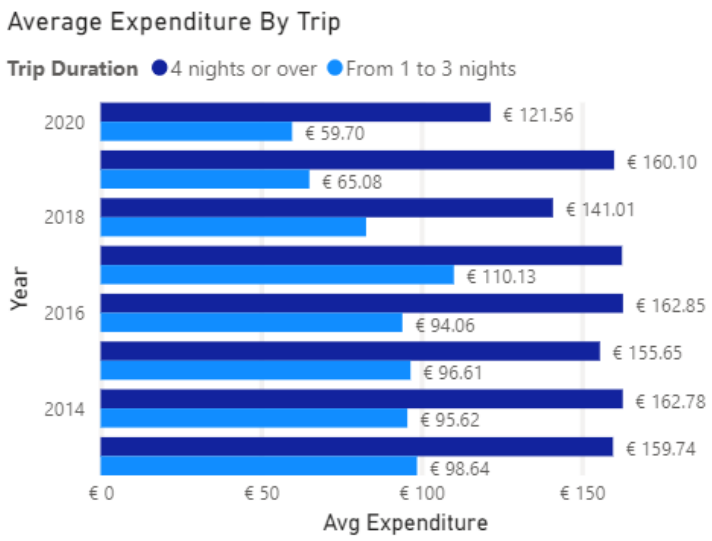
   Wartości w hurtowni:

	YearKey	(No column name)
1	2013	31.6043
2	2014	33.751
3	2015	32.0747
4	2016	31.9725
5	2017	35.03
6	2018	27.5695
7	2019	23.713
8	2020	18.5037

   Wyniki potwierdzają zgodność danych na wizualizacji.

3. Test wizualizacji: wykres słupkowy “Average Expenditure by Trip”

Wartości na wizualizacji:



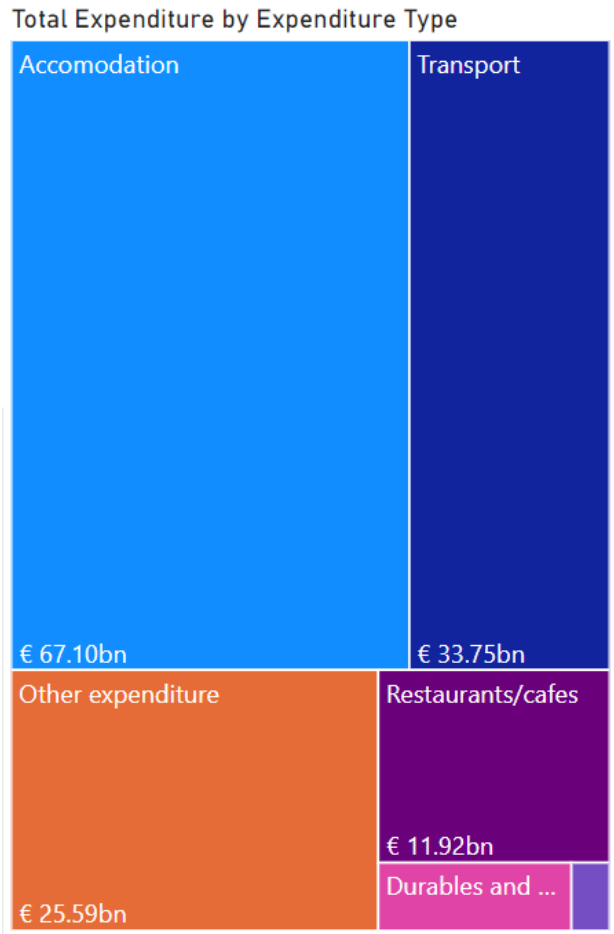
Wartości w hurtowni:

	YearKey	TripDuration	(No column name)
1	2020	4 nights or over	121.5586
2	2020	From 1 to 3 nights	59.695
3	2019	4 nights or over	160.0966
4	2019	From 1 to 3 nights	65.082
5	2018	4 nights or over	141.0113
6	2018	From 1 to 3 nights	82.763
7	2017	4 nights or over	162.5533
8	2017	From 1 to 3 nights	110.131
9	2016	4 nights or over	162.8545
10	2016	From 1 to 3 nights	94.0645
11	2015	4 nights or over	155.6487
12	2015	From 1 to 3 nights	96.6125
13	2014	4 nights or over	162.782
14	2014	From 1 to 3 nights	95.617
15	2013	4 nights or over	159.7391
16	2013	From 1 to 3 nights	98.6445

Wartości uzyskane na oba sposoby są identyczne, co oznacza, że wizualizacja działa poprawnie.

4. Test wizualizacji: wykres kafelkowy “Total Expenditure by Expenditure Type”

   Wartości na raporcie:



   Wartości w hurtowni:

	ExpenditureType	(No column name)
1	Accomodation	67096651563.13
2	Transport	33748776660.62
3	Other expenditure	25594460878.25
4	Restaurants/cafes	11924704924.84
5	Durables and valuable goods	3509426498.24
6	Package arrangement	694040919.90

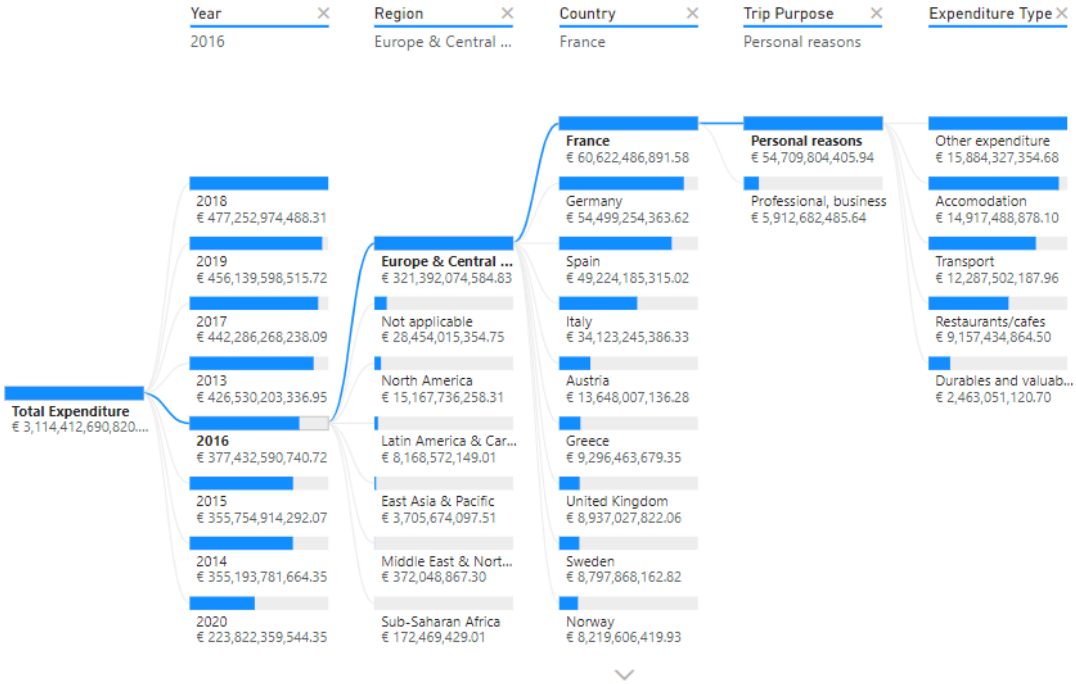
   Wyniki testu potwierdzają poprawność danych na wizualizacji.

Strona “Money flow”

W trakcie testów na tej stronie zastosowany był filtr ograniczający przedział lat od 2013 do 2020.

1. Test wizualizacji: Drzewo dekompozycji

   Wartości w raporcie:



Wyniki zapytań na hurtowni (kolejne zapytania dotyczą kolejnych warstw drzewa, zaczynając od lewej):

	TotalExpenditure
1	3114412690820.56

	YearKey	(No column name)
1	2018	477252974488.31
2	2019	456139598515.72
3	2017	442286268238.09
4	2013	426530203336.95
5	2016	377432590740.72
6	2015	355754914292.07
7	2014	355193781664.35
8	2020	223822359544.35

	Region	(No column name)
1	Europe & Central Asia	321392074584.83
2	Not applicable	28454015354.75
3	North America	15167736258.31
4	Latin America & Caribbean	8168572149.01
5	East Asia & Pacific	3705674097.51
6	Middle East & North Africa	372048867.30
7	Sub-Saharan Africa	172469429.01

	CountryName	(No column name)
1	France	60622486891.58
2	Germany	54499254363.62
3	Spain	49224185315.02
4	Italy	34123245386.33
5	Austria	13648007136.28
6	Greece	9296463679.35
7	United Kingdom	8937027822.06
8	Sweden	8797868162.82
9	Norway	8219606419.93

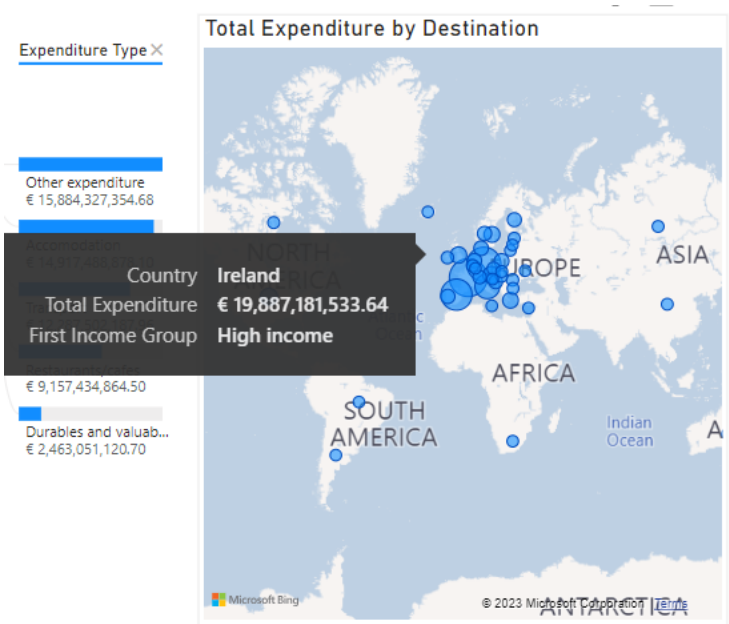
	TripPurpose	(No column name)
1	Personal reasons	54709804405.94
2	Professional, business	5912682485.64

	ExpenditureType	(No column name)
1	Other expenditure	15884327354.68
2	Accomodation	14917488878.10
3	Transport	12287502187.96
4	Restaurants/cafes	9157434864.50
5	Durables and valuable goods	2463051120.70

Wartości uzyskane poprzez kolejne zapytania SQL pokrywają się z tymi wyświetlanymi na drzewie dekompozycji co dowodzi jego poprawności.

2. Test wizualizacji: Mapa “Total Expenditure by Destination”

Przykładowa wartość na mapie:



Ta sama wartość uzyskana z hurtowni:

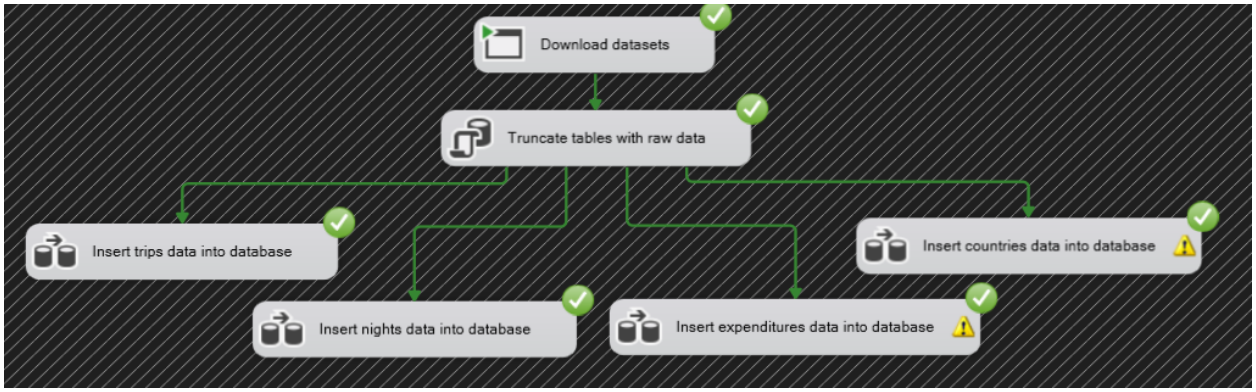
19	Ireland	High income	19887181533.64

Wyniki testu potwierdzają poprawność wizualizacji.

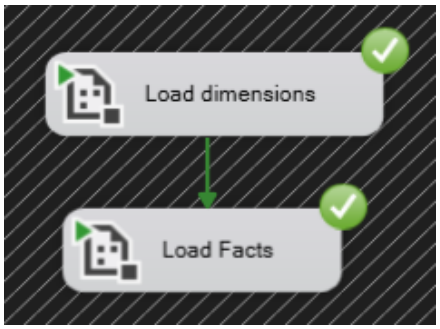
Testy end-to-end

Na sam koniec, w celu zweryfkowania działania systemu jako całości, od zera przeprowadzone zostały wszystkie kolejne etapy działania systemu. Oznacza to, że stworzona została od zera pusta struktura zarówno hurtowni jak i pośredniej bazy danych. Oczekiwanym rezultatem testu jest pomyśle działanie kolejnych komponentów od początku do końca, tj. pobrania danych, załadowania ich do hurtowni oraz na sam koniec wyświetlenia ich w narzędziu BI.

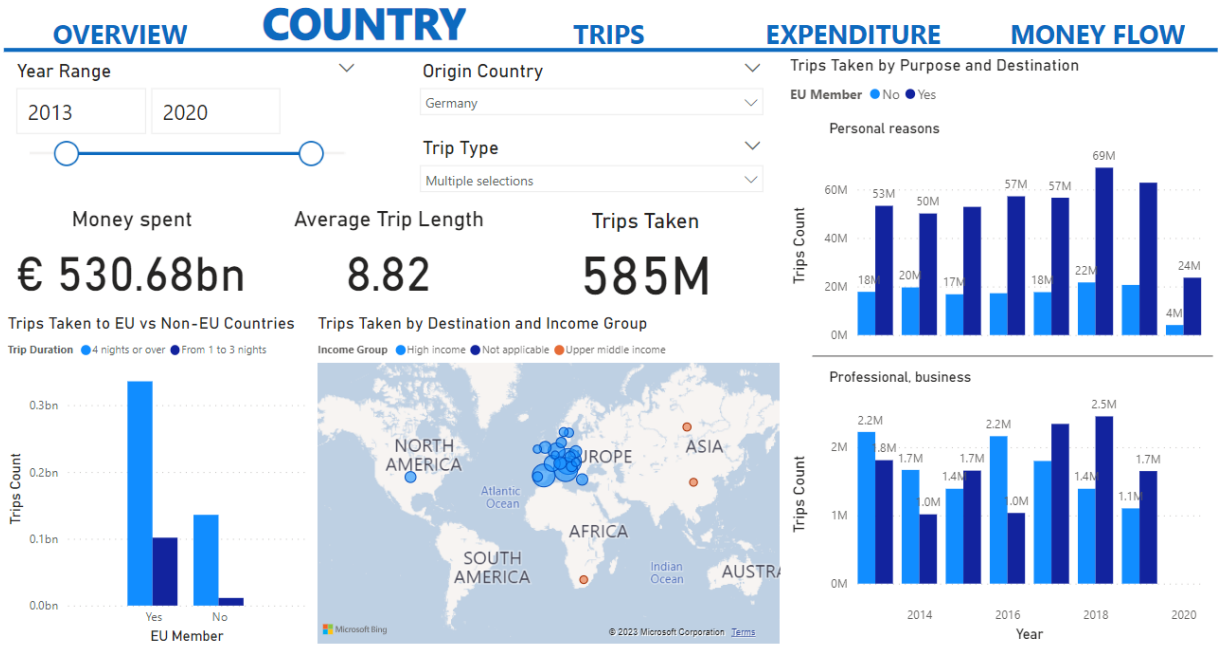
Pierwszym krokiem było uruchomenie procesu pobrania danych i załadowania ich do bazy pośredniej. Zgodnie z oczekiwaniami, proces ten zadziałał bez żadnych błędów:



Następnie uruchomiony został proces ETL w celu przekształcenia i załadowania danych do hurtowni. Ten etap również zadziałał bezbłędnie:



Na sam koniec uruchomione zostało Power BI, które w naszym przypadku zostało skonfigurowane tak, aby do obliczeń wykorzystywało serwer hurtowni poprzez wykorzystanie mechanizmu DirectQuery. Po uruchomieniu narzędzia oczekuje się zatem, że uprzednio załadowane dane będą dostępne do przeglądu przez użytkownika poprzez interfejs narzędzia.



W rzeczy samej, dane poprawnie wyświetlają się na każdej stronie raportu (wyżej wklejona została jedna ze stron).

Wyniki testu wskazują na to, że wszystkie komponenty zaprojektowanego przez nas systemu tworzą spójną i działającą całość.

Podsumowanie

Rozwiązanie dostarcza kompleksowych analiz i bogatych danych na temat trendów turystycznych w Unii Europejskiej, co przyczynia się do podejmowania trafnych decyzji biznesowych oraz skuteczniejszego zarządzania w sektorze turystycznym. Dostęp do szczegółowych informacji dotyczących podróży, wydatków i preferencji turystów, który umożliwia personalizację ofert, tworzenie skutecznych strategii marketingowych oraz odkrywanie nowych możliwości biznesowych.

Z perspektywy biznesu, rozwiązanie pozwala na identyfikację nowych segmentów rynku, nisz oraz potencjalnych obszarów rozwoju i innowacji. Umożliwia monitorowanie wyników działalności turystycznej, pomiar efektywności działań marketingowych oraz ciągłe doskonalenie strategii. W rezultacie firmy w sektorze turystycznym zyskują konkurencyjną przewagę, maksymalizują rentowność operacyjną i zwiększają satysfakcję klientów.

Z kolei dla rządu, rozwiązanie dostarcza cennych informacji, które mogą być wykorzystane do planowania strategicznego rozwoju turystyki w Unii Europejskiej. Rząd ma możliwość monitorowania i oceny efektywności działań, identyfikacji obszarów wymagających wsparcia oraz podejmowania odpowiednich decyzji politycznych. Analiza

danych umożliwia również lepsze zrozumienie wpływu sektora turystycznego na gospodarkę, co może prowadzić do tworzenia korzystnych warunków dla rozwoju branży oraz generowania większych dochodów.

Podsumowując, stworzenie analitycznych rozwiązań opartych na danych turystycznych przynosi liczne korzyści zarówno z perspektywy biznesu, jak i rządu. Jest to kluczowy element przyczyniający się do wzrostu konkurencyjności, efektywności operacyjnej i generowania większych korzyści dla wszystkich zainteresowanych stron.

Podział pracy w zespole

W ramach projektu, podział pracy w zespole został zrealizowany poprzez wspólne sesje pracy, w trakcie których wszyscy uczestnicy projektu aktywnie brali udział we wszystkich etapach. Praca była oparta na koncepcji współpracy i zaangażowania członków zespołu w każdym aspekcie projektu.

Podczas sesji pracy wspólnie analizowaliśmy wymagania projektu, zidentyfikowaliśmy cele oraz projektowaliśmy rozwiązania. Każdy z nas wnosił swoje unikalne spojrzenie i doświadczenie, co przyczyniało się do twórczego dialogu i skutecznej generacji pomysłów, usprawniało pracę oraz umoźliwiało szybkie rozwiązywanie powstających problemów. W trakcie spotkań, zadania były rozdzielane na bieżąco, w zależności od obszaru działania. Wspólnie realizowaliśmy poszczególne etapy projektu, dzieląc się obowiązkami i współpracując w celu osiągnięcia najlepszych rezultatów.