Hurtownie danych – Spr. 4.

PWr. WIZ, Data: 28-30.03.2022

Student	----------	Ocena
Indeks	256305	
Imię	Grzegorz	
Nazwisko	Dzikowski	

Zestaw składa się z 3 zadań. Jeżeli nie potrafisz rozwiązać zadania, to próbuj podać, chociaż częściowe rozwiązanie lub uzasadnienie przyczyny braku rozwiązania. Pamiętaj o podaniu nr. indeksu oraz imienia i nazwiska.

Baza danych: **AdventureWorks2017 lub 2019**

Zad. 1. - opcjonalne
Proszę przygotować dane do analizy zamówień w zakresie przedstawiony w tab. 1. (2 warianty rozwiązania) – (różnymi rozwiązaniami) – biorąc pod uwagę kryteria jakościowe – które z tych rozwiązań jest lepsze i dlaczego? Prezentacja graficzna z wykorzystaniem Tablau, Analiza i Wnioski graficzne - statystyki pamici, operacji, execution cost itp. itd

Tab. 1. Liczba zamówień w poszczególnych latach globalnie oraz obszarowo

Rok	Liczba zamówień	Terytorium	Liczba zam. na terytorium	% udział
2011	1607	Australia	463	28.81
2011	1607	Southwest	339	21.10
2011	1607	Germany	81	5.04
2011	1607	Central	50	3.11
2011	1607	Northwest	224	13.94

Rek.: **5/40**

Rozwiązanie:

Rozwiązanie 1

```
USE AdventureWorks2019;


WITH Orders_Territory_Years(Rok, [Total Orders], [Terytorium], [Ter
Orders], [rn])
AS
(
      SELECT
            YEAR(Sales.SalesOrderHeader.OrderDate),
            COUNT(Sales.SalesOrderHeader.SalesOrderID) OVER (PARTITION BY
Year(Sales.SalesOrderHeader.OrderDate)),
            Sales.SalesTerritory.[Name],
          COUNT(Sales.SalesOrderHeader.SalesOrderID) OVER (PARTITION BY
Year(Sales.SalesOrderHeader.OrderDate),
(Sales.SalesTerritory.TerritoryID)),
```

```
            ROW_NUMBER() OVER (PARTITION BY
Year(Sales.SalesOrderHeader.OrderDate), Sales.SalesTerritory.TerritoryID
ORDER BY Sales.SalesTerritory.[Name])
      FROM Sales.SalesOrderHeader
            JOIN Sales.SalesTerritory
                  ON Sales.SalesOrderHeader.TerritoryID =
Sales.SalesTerritory.TerritoryID


)

SELECT
      Rok,
      [Total Orders] as [Liczba zamówień],
      Terytorium,
      [Ter Orders] as [Liczba zam. na terytorium],
      FORMAT((CAST([Ter Orders] as FLOAT)/[Total Orders])*100, '##.##')
as [% udział]
FROM
      Orders_Territory_Years
WHERE rn = 1
ORDER BY
      Rok,
      Terytorium
```

Rok	Liczba zamówień	Terytorium	Liczba zam. na terytorium	% udział
2011	1607	Australia	463	28.81
2011	1607	Canada	149	9.27
2011	1607	Central	50	3.11
2011	1607	France	70	4.36
2011	1607	Germany	81	5.04

Roz: 5/40

Rozwiązanie 2

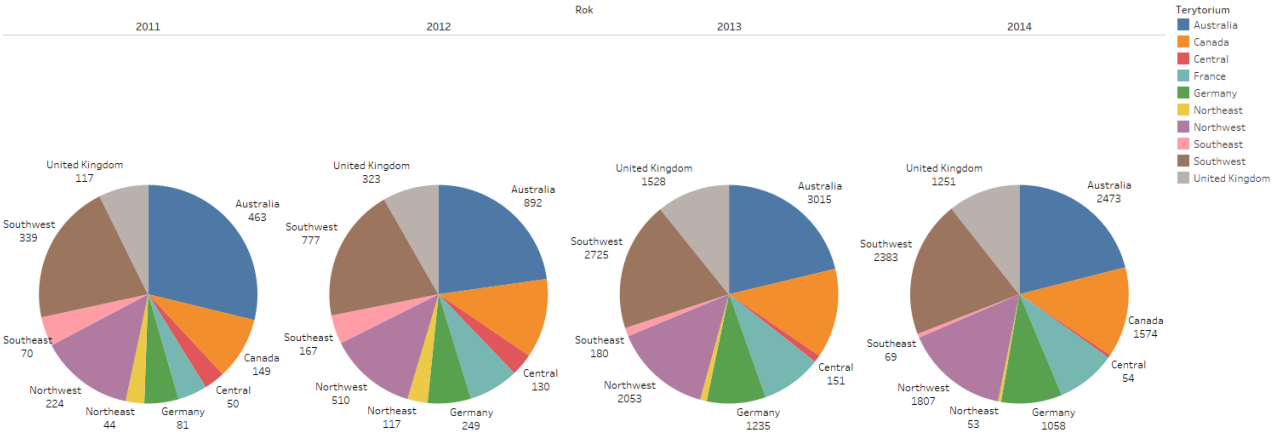
```
USE AdventureWorks2019;


WITH Orders_Territory_Years(Rok, [Orders], [Terytorium])
AS
(
      SELECT
            YEAR(Sales.SalesOrderHeader.OrderDate),
            COUNT(Sales.SalesOrderHeader.SalesOrderID),
            Sales.SalesTerritory.[Name]
      FROM Sales.SalesOrderHeader
            JOIN Sales.SalesTerritory
                  ON Sales.SalesOrderHeader.TerritoryID =
Sales.SalesTerritory.TerritoryID
```

```
	GROUP BY
		YEAR(Sales.SalesOrderHeader.OrderDate),
		Sales.SalesTerritory.[Name]

)


SELECT
	p.Rok,
	total.[Total Orders] as [Liczba zamówień],
	p.Terytorium,
	p.[Orders] as [Liczba zam. na terytorium],
	FORMAT((CAST(p.[Orders] as FLOAT)/total.[Total Orders])*100,
'##.##') as [% udział]
FROM
	Orders_Territory_Years p
	JOIN
	(
		SELECT
		Rok as Rok2,
		SUM(Orders) as [Total Orders]
		FROM Orders_Territory_Years
			GROUP BY Rok

	) total
	ON total.Rok2 = p.Rok
ORDER BY Rok, Terytorium;
```

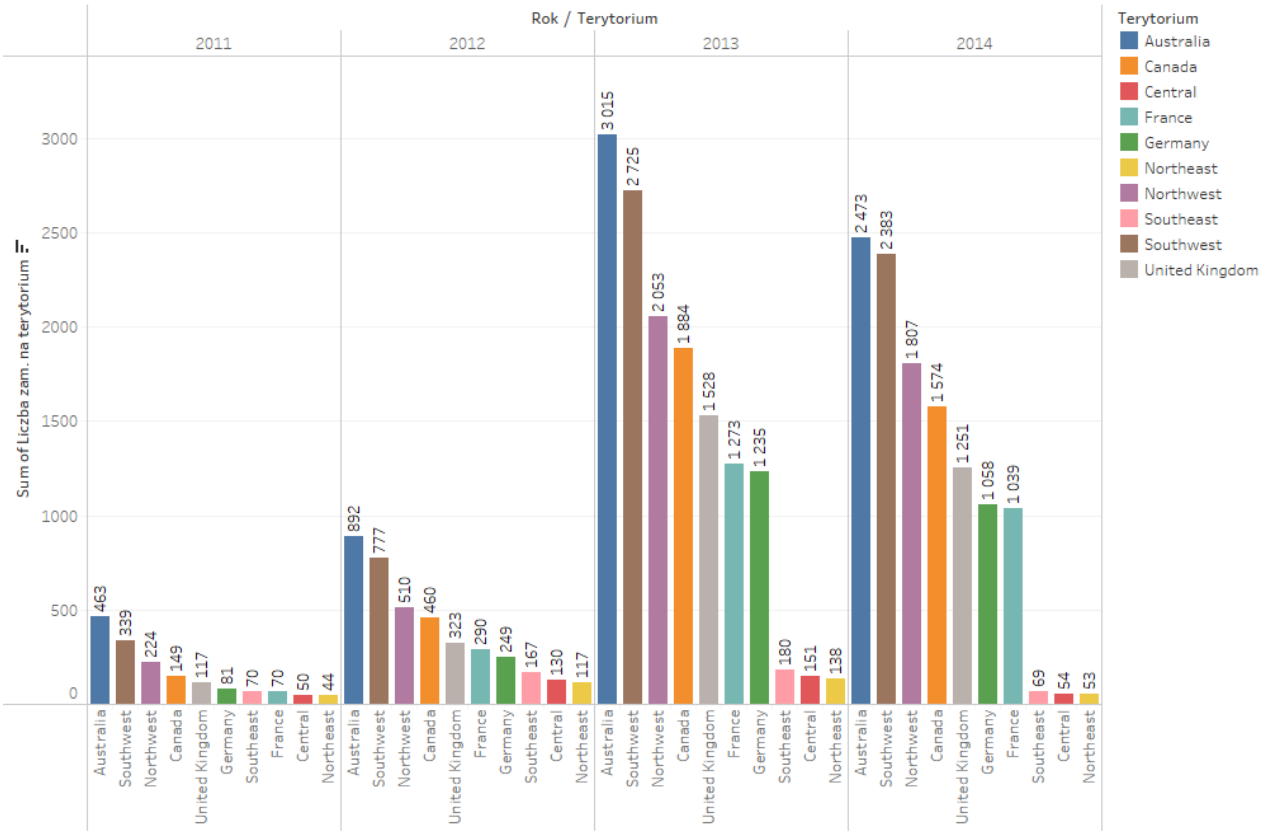
Rok	Liczba zamówień	Terytorium	Liczba zam. na terytorium	% udział
2011	1607	Australia	463	28.81
2011	1607	Canada	149	9.27
2011	1607	Central	50	3.11
2011	1607	France	70	4.36
2011	1607	Germany	81	5.04

Roz: 5/40

Udział poszczególnych terytoriów w rocznej sprzedaży produktów w latach 2011 - 2014



Roczna sprzedaż produktów w terytoriach w latach 2011 - 2014

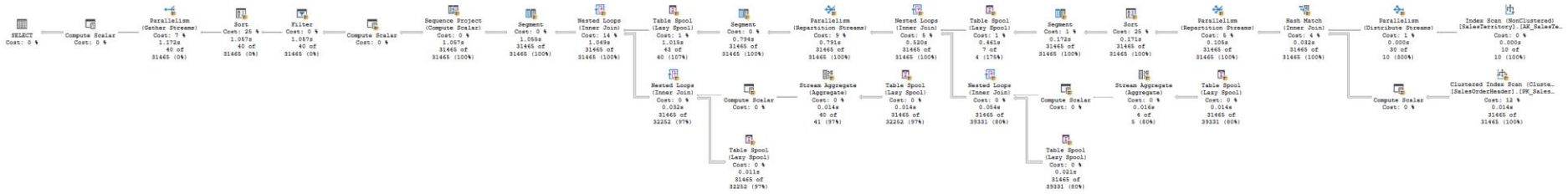


Analiza wydajności obu kwerend:

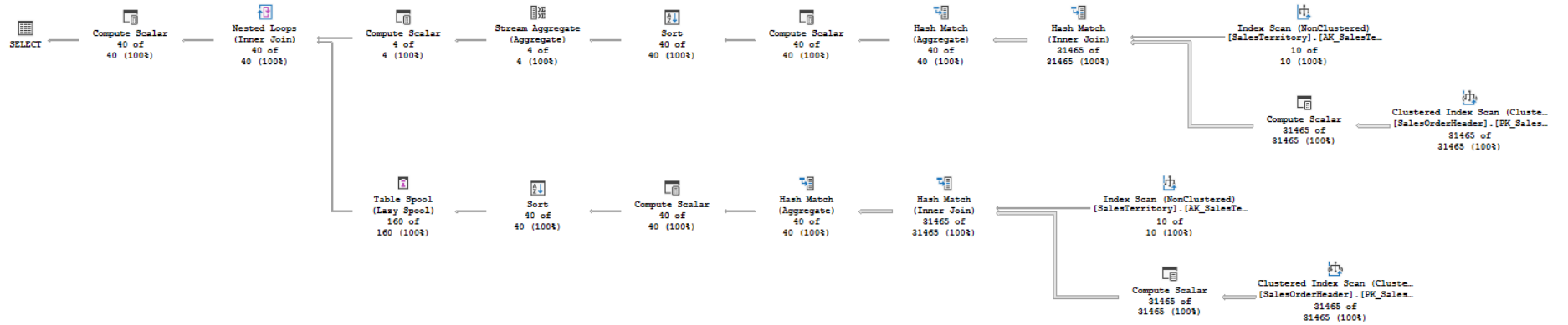
Rozwiązanie	CPU	Odczyty	Zapisy	Długość trwania
Rozwiązanie 1	1672	131880	408	1241
Rozwiązanie2	203	1472	0	200

Na następnej stronie znajdują się plany wykonania każdej kwerendy
Pierwsza moja myśl to było wykonanie kwerendy przy pomocy partition by. Wydaje się to rozwiązanie logiczne, ponieważ powala mi to zagregować poszczególne wartości w jednym rzędzie, a następnie zgrupować terytorium i rok w jeden wiersz. Jednak okazało się, że nie jest to optymalne rozwiązanie. Pętle, sortowania w partition by, oraz zrównoleglenie powoduje, że kod wykonuje się ponad 1 sekundę, wykonuje prawie 130 tysięcy odczytów i generalnie jest wolny
Drugie rozwiązanie zakłada użycie tylko i wyłącznie group by. Wynikowy SQL jest według mnie mniej czytelny, ponieważ wymaga użycia zagnieżdżonego SELECTa, jednak gdy zobaczymy na plan wykonania, to widać, że już na etapie pobrania danych są one zagregowane, tak więc dalsze przetwarzanie operuje tylko na ~40 wierszach tej kwerendy. Z tego powodu czas wykonania jest 6 krotnie szybszy, nie wykonuje żadnego zapisu oraz wymaga tylko ułamek tego, co pierwsza kwerenda.
Byłem zaskoczony tym wynikiem, ponieważ wydawało mi się, że group by będzie trwał długo, a partition wykona się szybko. Jednak na podstawie planów i czasów definitywnie wygrywa group by.

Rozwiązanie 1



Rozwiązanie 2



Zad. 2. Analiza danych i ocena ich jakości

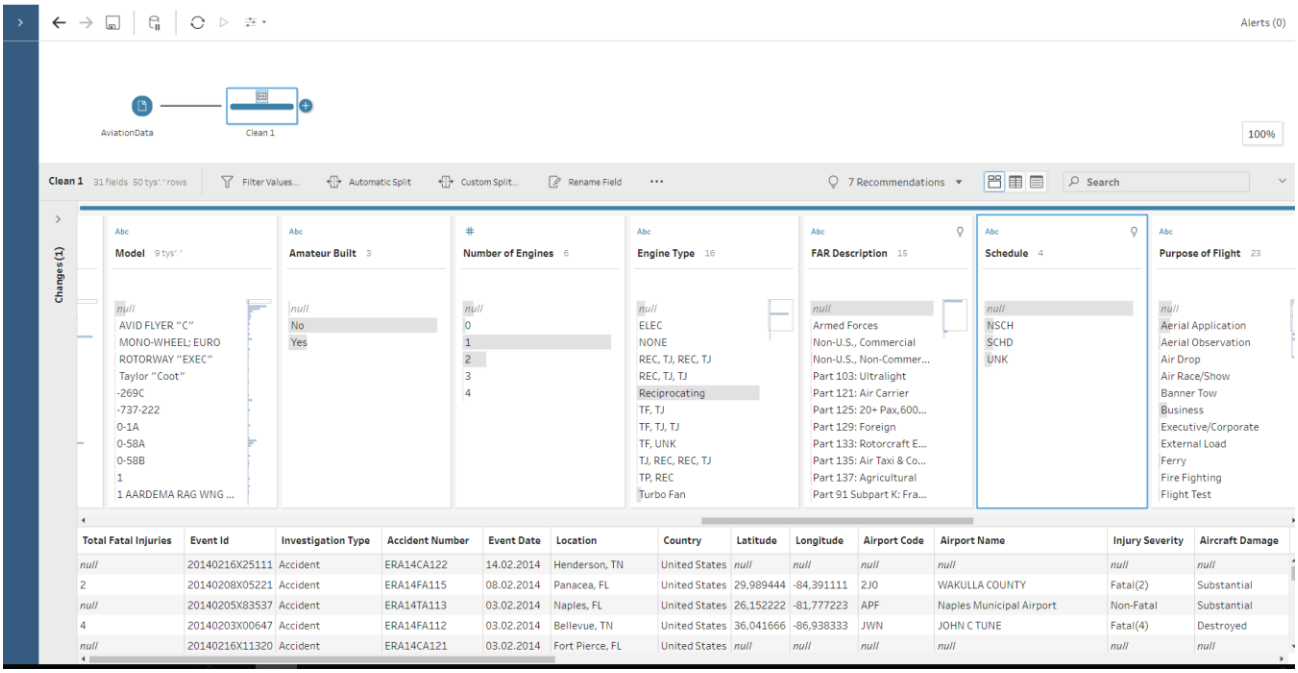
Przeanalizować, scharakteryzować i ocenić dane znajdujące się w pliku „**AviationData.xls**”, wykorzystując Tableau Prep oraz profilowanie danych z pakietu SSIS (projekt SQL Server Data Tools). Rozwiązanie przedstawić zgodnie z zakresem przedstawionym w tabelach 4.1. (słownik danych dziedzinowych) i 4.2 (ocena jakości danych źródłowych) dla przykładowego źródła danych (globalterrorism.csv).

Uwaga!

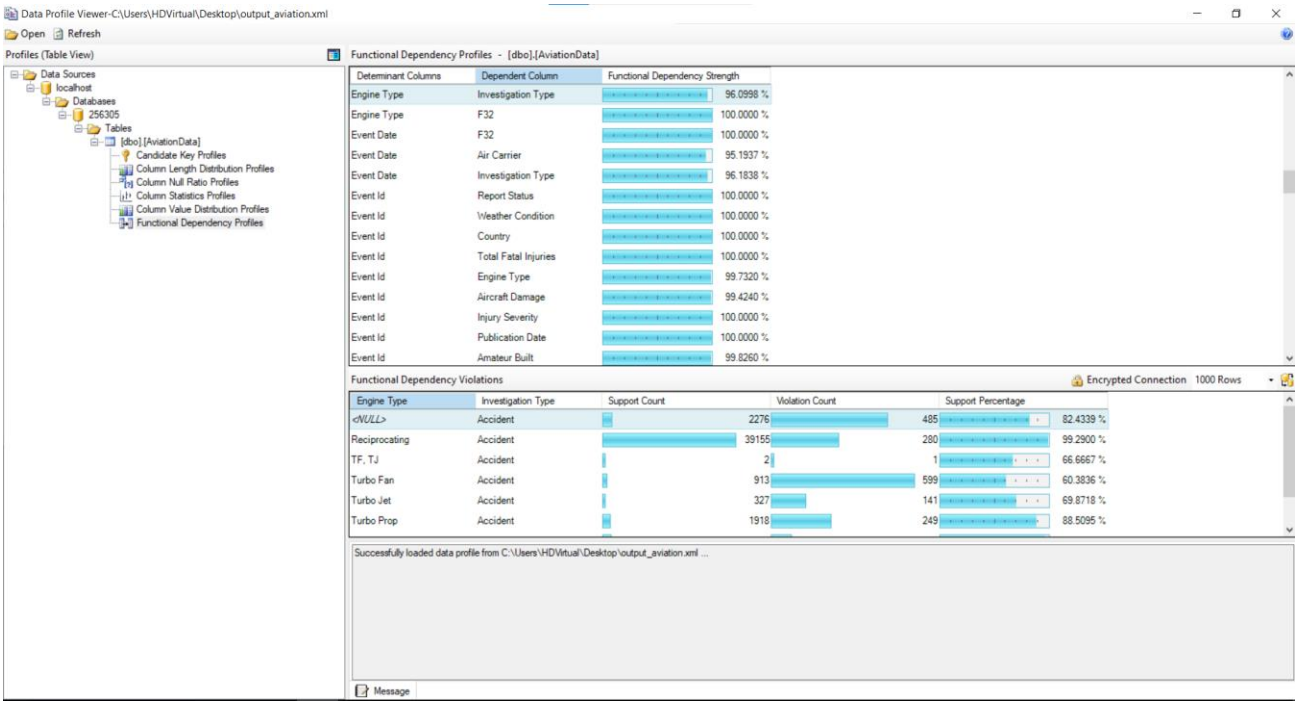
- **Analiza danych powinna być zrealizowana z wykorzystaniem pakietu Tableau Prep (materiały szkoleniowe można znaleźć na stronie Tableau oraz w Internecie)**
- Profilowanie danych z wykorzystaniem SSIS zostanie zrealizowane na zajęciach lab.

Rozwiązanie:

Kontekst: Zespół projektowy zlecił wykonanie raportu, który ma pokazać, jak na poważność wypadku wpływa typ pojazdu i odległość od miejsca wypadku



Rysunek 1 Analiza z wykorzystaniem Tablau Prep



Rysunek 2 Analiza z wykorzystaniem Profilera

Tabela 4.1 Interpretacja danych w kontekście rozważanej dziedziny problemowej (słownik danych)

Plik: aviationdata.csv			
Lp.	Atrybut	Typ wartości	Znaczenie
1.	EventId	Tekstowy	Kod identyfikujący incydent, podawany dla każdego incydentu
2.	InvestigationType	Tekstowy	Typ wydarzenia
3.	AccidentNumber	Tekstowy	Numer wydarzenia nadany przez Narodową Radę Bezpieczeństwa Transportu
4.	EventDate	Data	Data wystąpienia wydarzenia
5.	Location	Tekstowy	Przybliżona lokalizacja wystąpienia wydarzenia
6.	Country	Tekstowy	Kraj wystąpienia wydarzenia, null jeżeli nieznany
7.	Latitude	Numeryczny, zmiennoprzecinkowy	Szerokość geograficzna wystąpienia wydarzenia, null jeżeli nieznana
8.	Longitude	Numeryczny, zmiennoprzecinkowy	Długość geograficzna wystąpienia wydarzenia, null jeżeli nieznana
9.	AirportCode	Tekstowy	Kod Lotniska, z którego wyleciał pojazd. Null jeżeli nieznane

10.	AirportName	Tekstowy	Nazwa Lotniska, z którego wyleciał pojazd. Null jeżeli nieznane
11.	InjureSeverity	Tekstowy	Poważność zdarzenia
12.	AircraftDamage	Tekstowy	Jak poważnie został uszkodzony pojazd?
13.	AircraftCategory	Tekstowy	Kategoria pojazdu latającego.
14.	RegistrationNumber	Tekstowy	Numer boczny pojazdu
15.	Make	Tekstowy	Nazwa producenta pojazdu latającego
16.	Model	Tekstowy	Model pojazdu latającego
17.	AmateurBuilt	Prawda/Fałsz	Czy został wybudowany przez amatora?
18.	NumberOfEngines	Numeryczny, liczba całkowita	Ilość silników
19.	EngineType	Tekstowy	Typ silnika
20.	FARDescription	Tekstowy	Definicja według FAR (Federal Acquisition Regulation)
21.	Schedule	Tekstowy	Czy lot był zaplanowany?
22.	PurposeOfFlight	Tekstowy	Cel lotu
23.	AirCarrier	Tekstowy	Przewoźnik I właściciel pojazdu latającego
24.	TotalFatalInjuriesStat	Numeryczny, całkowity	Sumaryczna liczba ofiar śmiertelnych
25.	TotalSeriousInjuries	Numeryczny, całkowity	Sumaryczna liczba osób w ciężkim stanie
26.	TotalUninjured	Numeryczny, całkowity	Sumaryczna liczba osób bezobrażeń
27.	WeatherCondition	Tekstowy	Warunki pogodowe panujące w czasie wydarzenia
28.	BroadPhaseOfFlight	Tekstowy	Faza lotyu

29.	ReportStatus	Tekstowy	Status raportu
30.	PublicationDate	Data	Data publikacji raportu na temat wypadku

Tabela 4.2 Ocena jakości danych

Kontekst: Zespół projektowy zlecił wykonanie raportu, który ma pokazać, jak na poważność wypadku wpływa typ pojazdu i odległość od miejsca wypadku

Legenda:
Dane, które mają znaczenie dla analizy, i mają dobrą jakość
Dane, które mają znaczenie dla analizy, i mają średnią jakość
Dane, które mają znaczenie dla analizy, i mają słabą jakość
Dane nieistotne dla analizy

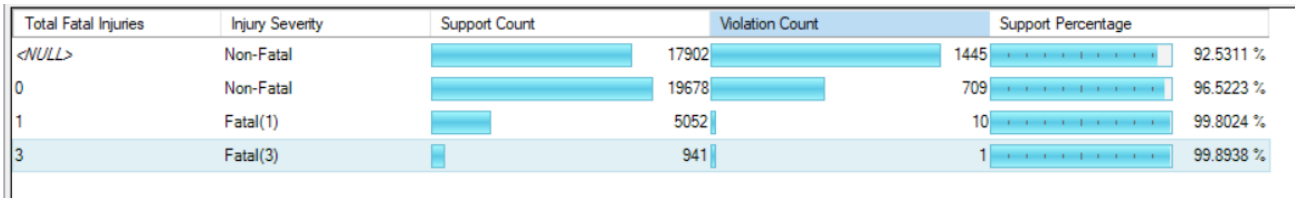
Plik: aviationdata.csv				
Lp.	Atrybut	Typ wartości	Zakres wartości	Ocena jakości danych
1.	EventId	Tekstowy	Dowolny unikalny tekst w formacie „YYYYMMDDXNNNNN” Gdzie YYYY-rok wydarzenia (1980-2020) MM – miesiąc wydarzenia (1-12) DD – dzień wydarzenia (1-31) NNNNN – unikalny numer wydarzenia	Dane identyfikujące konkretny wpis mają tylko 98.5% powiązania jako klucz, tak więc nie są one dobrym kluczem głównym 0% NULL
2.	Investigation Type	Tekstowy	Accident (Wypadek), Incident (Incydent) lub Null (Nieznany)	Praktycznie wszystkie dane mają określony typ wydarzenia, i te dane mają znaczenie w analizie, czy była to poważne wydarzenie czy nie <1% NULL (1 rekord)
3.	AccidentNumber	Tekstowy	Unikalny ciąg znaków, Długości od 9 do 11 znaków	Ten klucz ma 100% powiązania jako klucz, tak więc jest idealnym kandydatem jako klucz główny do analizy 0% NULL

4.	EventDate	Data	01.01.1980 – 01.01.2020	Dane dobrej jakości, wszystkie wydarzenia mają podaną datę, jednak data wydarzenia nie ma znaczenia dla tej analizy 0% NULL
5.	Location	Tekstowy	Istniejące miejsce na świecie (18964 unikalne wartości) – ciąg znaków długości od 4 do 61 znaków, NULL	Dane dobrej jakości, praktycznie wszystkie dane mają podaną lokalizację, i jest to istotne dla analizy, <1% NULL
6.	Country	Tekstowy	Ciąg znaków o długości od 4 do 30 znaków. Wszystkie unikalne wartości znajdują się w Tabeli 1. NULL	Dane dobrej jakości, mocno powiązany z lokalizacją (99.9% powiązania), istotne dla analizy <1% NULL
7.	Latitude	Numeryczny, zmiennoprzecinkowy	-78.016945 – 89.218056, NULL	Dana powiązana z Longitude, niestety większość danych nie posiada dokładnej lokalizacji, co czyni tę danę niskiej jakości 57% NULL
8.	Longitude	Numeryczny, zmiennoprzecinkowy	-193.21667 – 177.55778, NULL	Dana powiązana z Latitude, niestety większość danych nie posiada dokładnej lokalizacji, co czyni tę danę niskiej jakości Dodatkowo, jedna wartość jest niepoprawna w dziedzinie (długość geograficzna może mieć najmniej -180 stopni) 57% NULL
9.	AirportCode	Tekstowy	IATA Airport Code, o długości od 1 do 8 znaków, 7565 unikalnych wartości, NONE, NULL	42% NULL Aż 42% danych nie zawiera informacji o lokalizacji wylotowej, jednak dużo pojazdów na liście miało wylot z lokalizacji nieskodyfikowanej przez IATA. Ta dana powinna być mocno powiązana z AirportName, jednak 3% danych ma nazwę lotniska, ale nie ma jego kodu

10.	AirportName	Tekstowy	IATA Airport Name, 16354 unikalne wartości, o długości od 2 do 33 znaków, NULL	40% NULL Dane zawierają wartości N/A, None, NONE, które oznaczają brak danych (około 1%), oraz Private, PRIVATE, Private Airstrip, PRIVATE AIRSTRI, które oznaczają prywatne lotnisko (też około 1 %). Podobnie do AirportCode, nie wszystkie lokalizacje wylotowe mają nadany kod wyloty. 3% danych nie ma powiązania z AirportCode, bo nie wszystkie lokalizacje wylotowe mają nadany kod
11.	InjureSeverity	Tekstowy	Non-Fatal, Incident Unavailable, null lub Fatal. W przypadku Fatal w nawiasie podana będzie ilość ofiar śmiertelnych, np. Fatal(10) – 10 ofiar śmiertelnych. Zakres od 1 do 350	Większość danych ma dane o ofiarach, dana o wysokiej jakości <1% NULL
12.	AircraftDamage	Tekstowy	Destroyed, Minor, Substantial, NULL	Dobra jakość danych, większość danych ma podane stopień uszkodzenia pojazdu, 3% NULL
13.	AircraftCategory	Tekstowy	Airplane, Balloon, Blimp, Glider, Gyrocraft, Helicopter, Powered-Lift, Ultralight, Unknown, NULL	79% NULL Dane niskiej jakości, ponieważ prawie 80% danych nie jest podanych i przypisanych
14.	RegistrationNumber	Tekstowy	Unikalny numer boczny pojazdu, ciąg o długości od 3 do 11 znaków NULL	Dane bezpośrednio nie przydatne, jednak pośrednio, posiadając bazę danych zarejestrowanych pojazdów latających, możnaby dowiedzieć się o typie pojazdu, a dane są lepszej jakości od AircraftCategory 5% NULL
15.	Make	Tekstowy	Istniejący producent, ciąg znaków o długości 2 do 30 znaków NULL	Dana nie istotna w analizie <1% NULL

16.	Model	Tekstowy	Model pojazdu, ciąg znaków o długości od 1 do 20 znaków, NULL	Dana nie istotna w analizie <1% NULL
17.	AmateurBuilt	Prawda/Fałsz	Yes, No, NULL	Informacja, czy pojazd został wybudowany samodzielnie jest ważna, Dana dobrej jakości, istotna w analizie, 1% NULL
18.	NumberOfEngines	Numeryczny, liczba całkowita	0, 1, 2, 3, 4, NULL	Ilość silników to jedna z cech typu pojazdu, dana istotna w analizie i dobrej jakości, 6% NULL
19.	EngineType	Tekstowy	15 unikalnych wartości, ciągi znaków o długości od 4 do 16, NULL	Typ silnika to jedna z cech typu pojazdu, dana istotna w analizie i dobrej jakości 6% NULL
20.	FARDescription	Tekstowy	15 unikalnych wartości, NULL	Dana niskiej jakości, nie zawsze FAR posiada ścisłą definicję, a w przypadku tej danej NULL oznacza niesprecyzowaną definicję, 79% NULL
21.	Schedule	Tekstowy	NSCH, SCHD, UNK, NULL	Dana nieistotna w analizie, 86% NULL
22.	PurposeOfFlight	Tekstowy	22 unikalne wartości, ciągi znaków o długości 4 do 19 NULL	Dana nieistotna w analizie 6% NULL
23.	Air Carrier	Tekstowy	1825 unikalnych wartości, ciągi znaków o długości od 3 do 90, NULL	Dana nieistotna w analizie 95% NULL
24.	TotalFatalInjuries	Numeryczny, całkowity	0 – 349, NULL	Dana istotna w analizie, Null w tym przypadku może oznaczać kilka rzeczy. Niektóre dane w momencie, gdy Severity = Non-fatal, mają albo wartość 0, albo null. Dodatkowo, niektóre dane są niespójne – występuje 11 przypadków, gdzie ta dana jest niezgodna z Severity (Rysunek 3) 39% NULL (często może nie dotyczyc)

25.	TotalSeriousInjuries	Numeryczny, całkowity	0-111. NULL	Dana istotna w analizie, w tym przypadku NULL ma znaczenie podobne do 0, 42% NULL
26.	TotalUninjured	Numeryczny, całkowity	0-699, NULL	Dana istotna w analizie, w tym przypadku NULL ma znaczenie podobne do 0 40% NULL
27.	WeatherCondition	Tekstowy	IMC, UNK, VMC, null	Dana nieistotna w analizie 3% NULL
28.	BroadPhaseOfFlight	Tekstowy	12 unikalnych wartości, podanych w tabeli 2, NULL	Dana nieistotna w analizie 12% NULL + 0.612% UNKNOWN
29.	ReportStatus	Tekstowy	Factual, Foreign, Preliminary, Probable Cause	Dana nieistotna w analizie 0% NULL
39.	PublicationDate	Data	01.01.1980 – 01.01.2020, NULL	Dana nieistotna w analizie <1% NULL



Total Fatal Injuries	Injury Severity	Support Count	Violation Count	Support Percentage
<NULL>	Non-Fatal	17902	1445	92.5311 %
0	Non-Fatal	19678	709	96.5223 %
1	Fatal(1)	5052	10	99.8024 %
3	Fatal(3)	941	1	99.8938 %

Rysunek 3 Fatal Injuries and Injury Severity

Wniosek:
Analiza poważności wypadku na podstawie typu pojazdu i odległości od miejsca wypadku byłaby niewiarygodna na tym zestawie danych. Większość danych nie posiada dokładnych lokalizacji wypadku, oraz lotniska wylotowego, co mocno redukuje próbkę danych. Dodatkowo, niektóre dane są niepoprawne i można wywnioskować, że były wprowadzane ręcznie, a nie generowane.

Zad. 3. Analiza i wybór obszaru tematycznego (dziedziny problemowej) oraz propozycja tematu mini projektu hurtowni danych

Pomysłów mam bardzo dużo, jednak problemem są dane. Pierwszy pomysł, który dość mnie interesuje, to sprzęt komputerowy - temat związany z popytem na karty graficzne, jego wpływ na ceny, lub analiza sprzedażowa kart graficznych. Jednak po poszukiwaniu internetu nie możliwe jest znaleźć dane, zwłaszcza przynajmniej 50 tysięcy rekordów.

Drugim tematem była giełda i/lub rynek krypto, i tutaj są obiecujące dane historyczne, ale są ukryte za api, do którego trzeba dostać dostęp

Allegro posiada API, które wydawało się cenne, ale także jest mało szczegółowe. Dane – może udałoby się zdobyć, ale nie da się z nich wyodrębnić żadnych faktów, potrzebnych w przypadku hurtowni danych

Próbowałem także poszukać wycieków z firm, które są publiczne. Na przykład w 2020 roku wyciekła baza LinkedIn, która mogłaby być ciekawym punktem. Ale po pierwsze, nie jestem pewien, czy to jest legalne, a po drugie, moja moralność nie pozwala mi korzystać z wyciekniętych danych do analiz.

Finalnie zacząłem szukać danych związanych z grami. Jeden temat, który zaproponuje niżej, dotyczy 2b2t – serwera minecraft, który działa od 2011 roku. Ostatnio gracze stworzyli publiczny zapis mapy, który mógłby posłużyć pod analizę, wraz z danymi historycznymi sprzed kliku lat

Finalnie zacząłem poszukiwać tematu na podstawie dostępnych danych, znalazłem kilka stron i publicznych źródeł, i tak

3.1. Temat: Analiza rozpraw sądowych w sądzie najwyższym USA

3.2. Uzasadnienie: Znalazłem dane (120 tysięcy rekordów), które mogą posłużyć do analizy spraw sądowych w sądzie najwyższym USA. Dane są uporządkowane i można próbować je analizować pod kątem jednomyślności, jak sprawa się potoczyła, powodu itp. http://scdb.wustl.edu/data.php?s=1

3.3. Temat: Analiza giełdy

3.4. Uzasadnienie: Istnieje wiele stron, na których można za darmo lub za api pobrać dane historyczne o cenach akcji, do tego można zrobić przypisanie firm oraz ich działalność, co pozwala na dość dużą analizę danych na ten temat https://www.quantshare.com/sa-620-10-new-ways-to-download-historical-stock-quotes-for-free

3.5. Temat: Analiza danych dotyczących graczy 2b2t

3.6. Uzasadnienie: 2B2T to serwer minecrafta typu anarchia (brak jakichkolwiek zasad), który działa od 2011 roku. Są dostępne rozległe zapisy mapy z 2018, 2019, 2020 i 2021 roku. Na ich podstawie można analizować np. co było napisane w przedmiotach w danym roku, jak zmieniał się region startowy, przeprowadzić analizę kiedy i jak zostały wygenerowane nowe tereny, najczęściej odwiedzane miejsca itp. Nie jest to bardzo biznesowa analiza, ale interesuje się tą grą, dlatego pomyślałem, że przy takiej trudności znalezieniu danych spróbuję i tutaj.

Wnioski:

Pierwsze zadanie ciekawie pokazało, że SQL group by jest bardzo optymalny w porównaniu do partition by. Być może gdybym popracował nad optymalizacją, to dałoby się wyłuskać, ale group by dzięki swojemu szybkiej redukcji ilości wierszy bardzo przyśpiesza działanie.

Drugie zadanie uświadomiło mi, jak trudna jest ocena jakości danych. Siedziałem nad tym kilka godzin, analizując dostępne dane w Tablau Prep i Profile Viewerem. Jednak jestem zadowolony, bo udało mi się określić, że kontekst analizy, którą chciałem przeprowadzić w hipotetycznym przypadku, byłby niewiarygodny.

Trzecie zadanie uświadomiło mi, jak ciężko znaleźć dane do hurtowni, i nie dziwię się, że wielkie firmy, posiadające dane, nie chcą się nimi dzielić, skoro tak ciężko je zebrać. A wiarygodne i dobre dane są podstawą działania dobrych firm, więc jest to istotna sprawa.

Uwaga:

- Sprawozdanie bez wniosków końcowych nie będzie sprawdzane i tym samym ocena jest negatywna!

Zad. 3. (Prezentacja wszystkich punktów tego zadania na zajęciach 11-12.04.2022)

Proces tworzenia hurtowni danych powinien być poprzedzony zrozumieniem „potrzeb biznesu" oraz rzeczywistości (dziedziny problemowej) reprezentowanej przez dostępne zasoby danych. Realizacja poniższego zadania ma uzmysłowić występujące problemy w określonym (wybranym) wycinku rzeczywistości, a następnie umożliwić zidentyfikowanie (określenie) potrzeb, celu i możliwości analiz biznesowych, by wspierać procesy decyzyjne (podejmowanie właściwych decyzji biznesowych).

Projekt HD – propozycja tematu
Proszę przygotować zakres realizacji projektu zgodnie z poniższą specyfikacją oraz przedyskutować propozycję projektu z osobą prowadzącą zajęcia. Poczynione uzgodnienia zarejestrować w formie wniosków. Na zajęciach laboratoryjnych należy przedstawić na forum grupy swoją propozycję tematu projektu (uzasadniając celowość i jego główne elementy 1.1 – 1.6) wykorzystując PowerPoint.

Zakres opracowania projektu HD – cz. 1.

1.1. Tytuł projektu

1.2. Charakterystyka dziedziny problemowej

1.3. Opis obszaru analizy wraz z uzasadnieniem (wybrany fragment dziedziny, przeznaczony do szczegółowej analizy i opracowania hurtowni danych)

1.4. Problemy

1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania i potrzeby w zakresie wsparcia podejmowania decyzji

1.5.2. Zakres analizy – badane aspekty

1.6. Źródła danych

1.6.1. Lokalizacja, format, dostępność

1.6.2. Wstępna ocena (liczba rekordów, zakres czasowy danych - faktów)

1.6.3. Fakty

Lp.	Fakt	Miary
1.		
2.		
…	…	

1.6.4. Kontekst analizy faktów np. czas (ziarnistość), lokalizacja, warunki pogodowe, itd.

Lp.	Kontekst analizy - wymiary	Własności
1.		
2.		
3.		
…	…	

ZAŁĄCZNIK: TABELE

Country
Argentina
Australia
Bahamas
Brazil
Canada
Colombia
Ecuador
France
Germany
Indonesia
Italy
Japan
Mexico
Peru
Spain
United Kingdom
United States
Venezuela

Tabela 1 Wszystkie wartości kolumny country

BroadPhaseOfFlight
APPROACH
CLIMB
CRUISE
DESCENT
GO-AROUND
LANDING
MANEUVERING
OTHER

BroadPhaseOfFlight
STANDING
TAKEOFF
TAXI
UNKNOWN

Tabela 2 Wszystkie wartości kolumny BroadPhaseOfFlight