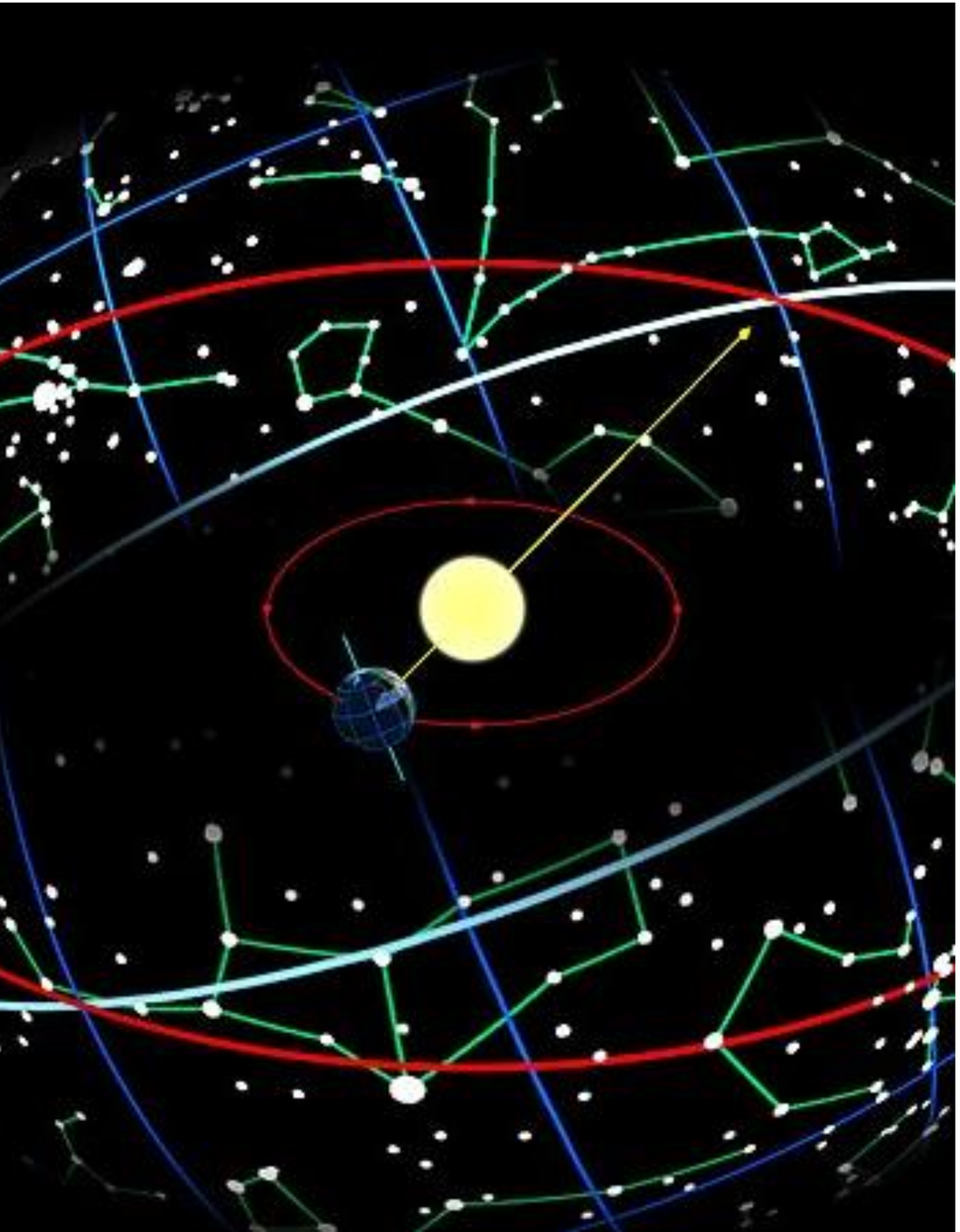
UNIWERSYTET
WSB MERITO
POZNAŃ

UFO - Obserwowalne zjawiska nadprzyrodzone na świecie w ciągu ostatniego stulecia

Bartosz Augustyniak - Lider
Waldemar Grzelka
Bartosz Kryspin
Mateusz Makowski
Mateusz Marecki

Poznań 2025

Spis treści



Opis projektu i cel analizy

🔍 Opis projektu

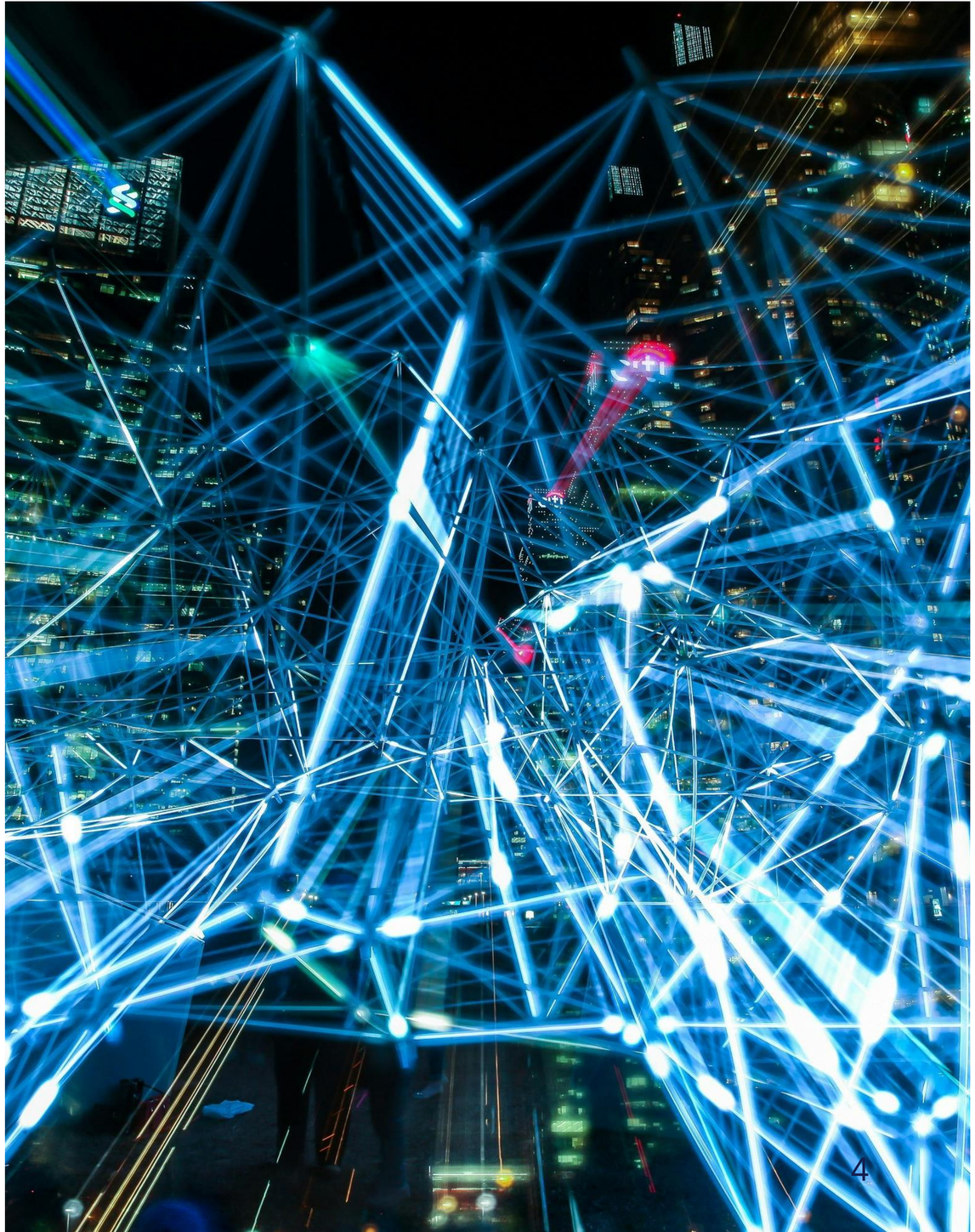
- **Zakres danych:** Ponad 80 000 zgłoszeń obserwacji UFO z ostatniego stulecia.
- **Źródło:** Publiczny zbiór danych dokumentujący miejsce, czas, opis i czas trwania obserwacji.
- **Wersje danych:**
  - **Pełna:** zawiera wszystkie rekordy, także z brakami lub błędami.
  - **Oczyszczona: przygotowana do analizy po usunięciu błędów i braków.**
- **Problemy danych:**
  - Ok. 12% zgłoszeń bez kompletnej lokalizacji (dane wejściowe).
  - Starsze dane są mniej kompletne ze względu na sposób rejestrowania**.**

🎯 Cel analizy

- **Identyfikacja trendów czasowych:** analiza zmian liczby zgłoszeń w czasie (np. dekady, pory roku).
- **Analiza przestrzenna:** wykrywanie obszarów z największą liczbą obserwacji (miasta, stany).
- **Wyszukiwanie korelacji:** zależności między czasem, lokalizacją i czasem trwania zjawisk.
- **Wizualizacja wyników w Power BI**

Główne etapy prac

1. Pozyskiwanie danych
•Identyfikacja źródeł danych: pliki CSV, API, bazy danych, dane strumieniowe, web scraping.
•Import danych do środowiska analitycznego.
2. Eksploracja i wstępna analiza
•Zrozumienie struktury danych: liczba rekordów, typy kolumn, zakresy wartości.
•Identyfikacja braków danych, błędów i anomalii.
3. Czyszczenie i przygotowanie danych
•Uzupełnianie lub usuwanie braków danych.
•Standaryzacja formatów (np. daty, jednostki).
•Usuwanie duplikatów, korekta błędów logicznych.
•Kodowanie danych.
4. Transformacja i wzbogacanie danych
•Tworzenie nowych zmiennych.
•Agregacje, grupowanie, przekształcenia czasowe.
•Łączenie danych z różnych źródeł.
5. Modelowanie i analiza danych
•Budowa modeli statystycznych i/lub ML do analizy trendów, klasyfikacji, predykcji.
•Analiza korelacji i zależności między zmiennymi.
6. Wizualizacja i raportowanie
•Tworzenie dashboardów w Power BI
•Interaktywne mapy, wykresy czasowe, heatmapy, tabele przestawne.



Oceny jakości danych źródłowych

📈 **Stan danych przed czyszczeniem:**
•Większość kolumn zawierała **ponad 99% danych niepustych (not null)**.
•**Wyjątki:**
  •state – 93%
  •country – 88%
  •shape – 98%

🧩 **Główne wyzwania:**
•Braki w danych lokalizacyjnych (state, country).
•Konieczność rekonstrukcji brakujących informacji na podstawie innych kolumn.

🧭 **Metody uzupełniania danych:**
•**Klucz łączony:** latitude + longitude – porównywanie lokalizacji z istniejącymi rekordami.
•**Ekstrakcja danych z kolumny city:**
  •Wydobycie fragmentu tekstu po znaku specjalnym (np. „-”, „(”).
  •Dopasowanie do listy krajów świata w celu eliminacji błędów.

📜 **Efekty działań:**
•**Kolumna country:** poprawa jakości do **96% wartości niepustych**.
•**Kolumna state:** brak istotnej zmiany (ograniczenia w dostępnych danych).

🖼 **Dokumentacja postępów:**
•Zrzuty ekranu pokazujące jakość i rozkład danych po oczyszczeniu:
  •Data_Cleansing_Sample_View_Complete.png
  •Data_Cleansing_Sample_View_Scrubbed.png

Przygotowanie i oczyszczanie danych – Power Query (Excel)

✅ **Dlaczego Power Query?**

•Dane wejściowe < 100 tys. wierszy – brak konieczności wykorzystania bardziej złożonych narzędzi;
•Intuicyjne zarządzanie transformacjami i ładowaniem danych.
•Możliwość stworzenia prostego pipeline'u ETL;
•Wbudowane narzędzia analizy danych: jakość, statystyki, rozkład.

🔄 **Etapy przetwarzania danych:**

1.Załadowanie danych źródłowych do Power Query;
2.Wyodrębnienie i indywidualna obróbka kolumn;
3.Scalanie kolumn na podstawie sztucznego klucza (ID wiersza);
4.Uzupełnianie braków w danych (głównie kolumny „state” i „country”).

💾 **Efekt końcowy:**

•Oczyszczony i ujednolicony zbiór danych zapisany do formatu CSV;
•Gotowość do załadowania do bazy danych SQL w chmurze **Microsoft Azure**.

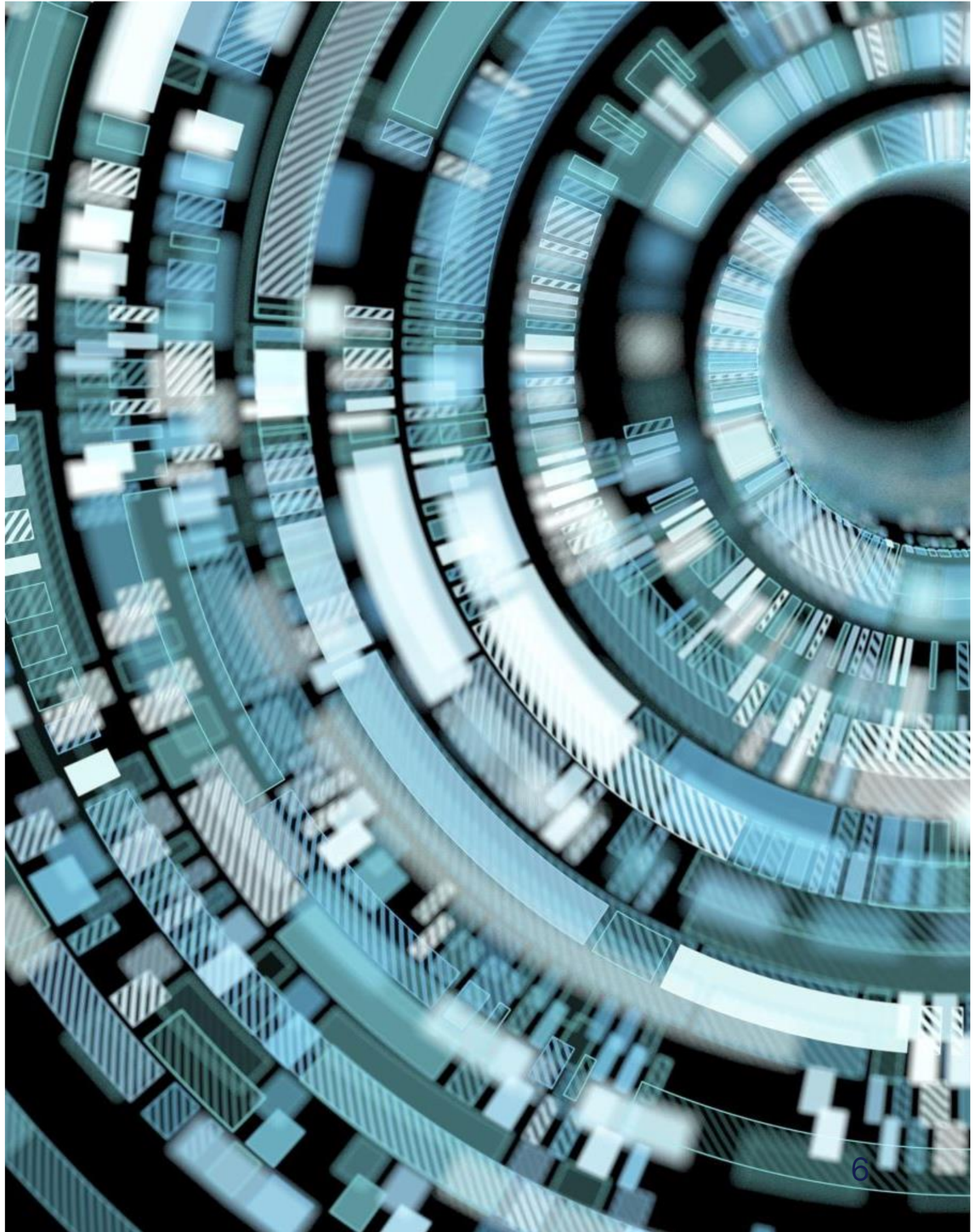
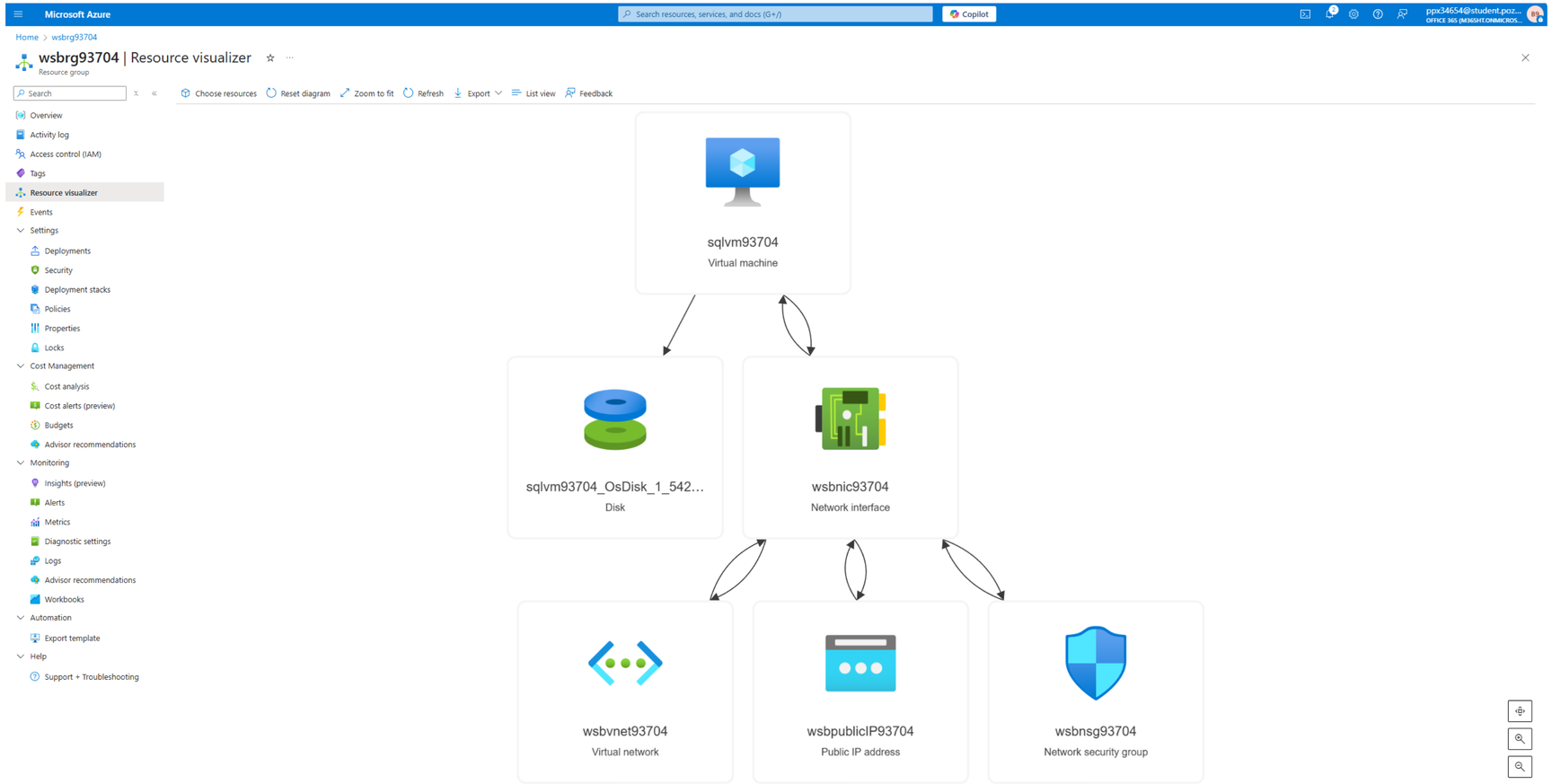


Diagram architektury oraz rozwiązanie chmurowe- Azure

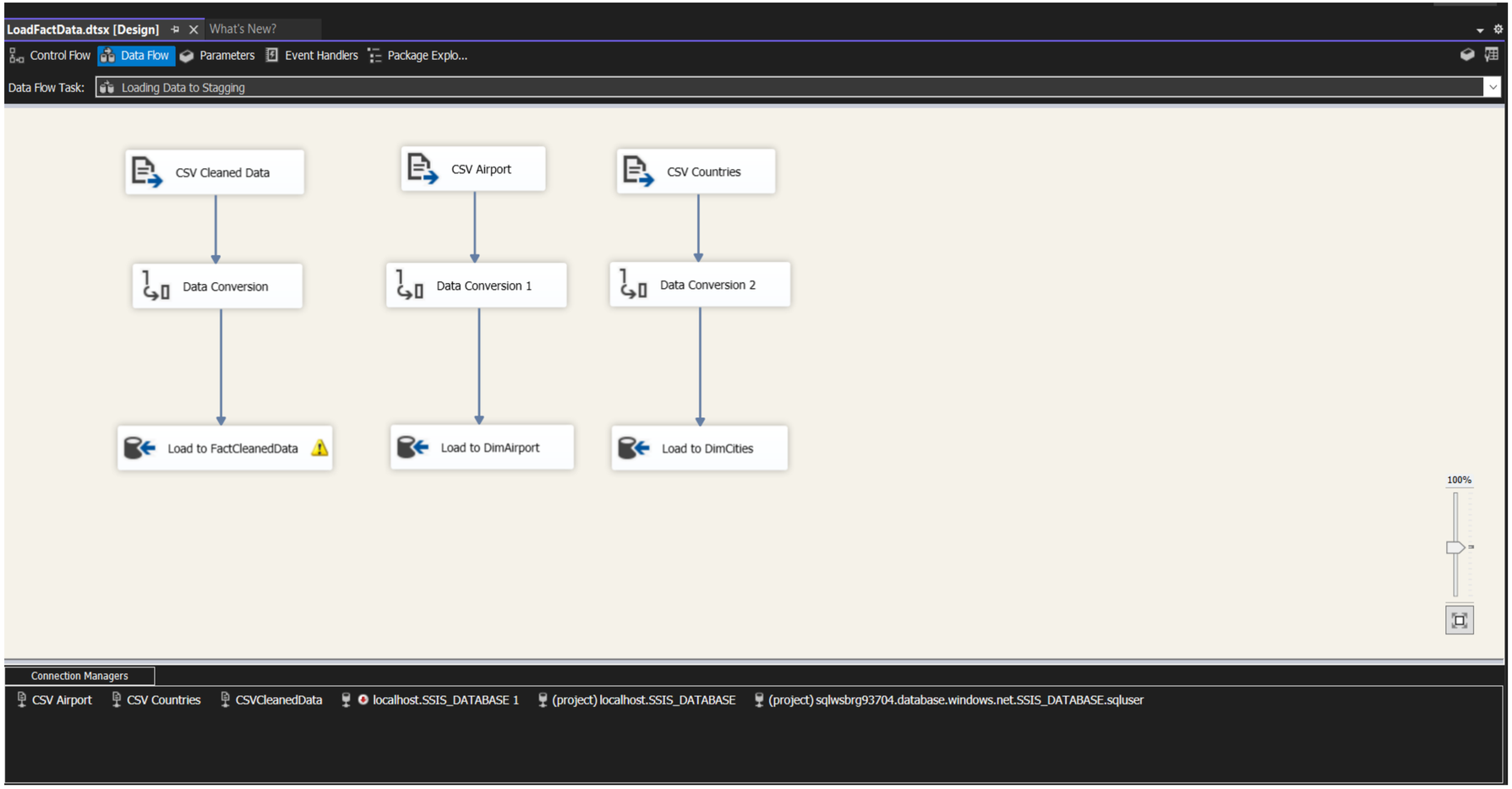


Azure Resource Visualizer

Schemat zasobów dla grupy „wsbrg93704”, uruchomionych na platformie Azure obejmujący:

- maszynę wirtualną,
- dysk,
- interfejs sieciowy,
- grupę zabezpieczeń sieci,
- wirtualną sieć oraz publiczny adres IP.

Diagram architektury oraz rozwiązanie chmurowe- Azure



SSIS Data Flow Task

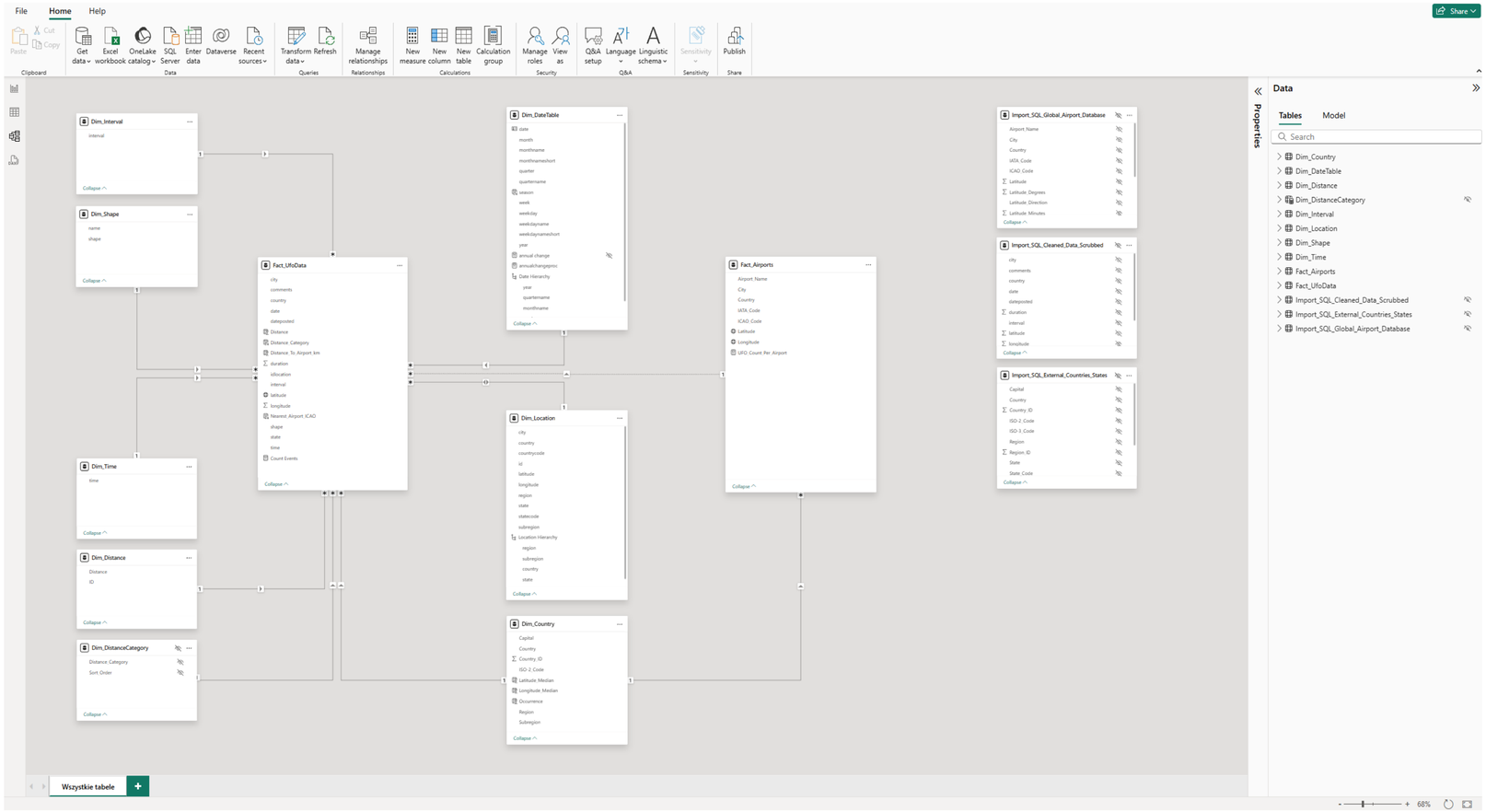
- SSIS Data Flow Task odpowiedzialny za ładowanie danych z pliku CSV do odpowiednich tabel w bazie danych sqlwsbrg93704/SSIS_DATABASE na platformie Azure.
- W tym tasku dla każdego pliku ładujemy dane do odpowiadających im tabel.
- W bazie danych sqlwsbrg93704/SSIS_DATABASE utworzono trzy tabele (jedna tabela faktów oraz dwie tabele wymiarów).

Diagram architektury oraz rozwiązanie chmurowe- Azure

- Paczka SSIS odpowiadająca całemu mechanizmu ładowania danych do tabel w bazie danych sqlwsbrg93704/SSIS_DATABASE na platformie Azure.
- W paczce jest SQL Task odpowiadające za usuwanie danych z poszczególnych tabel (TRUNCATE TABLE) oraz Data Flow Task Loading Data to Stagging, odpowiadający załadowaniu pliku z plikami CSV do tabel w bazie danych sqlwsbrg93704/SSIS_DATABASE.

SSIS Load Data Package

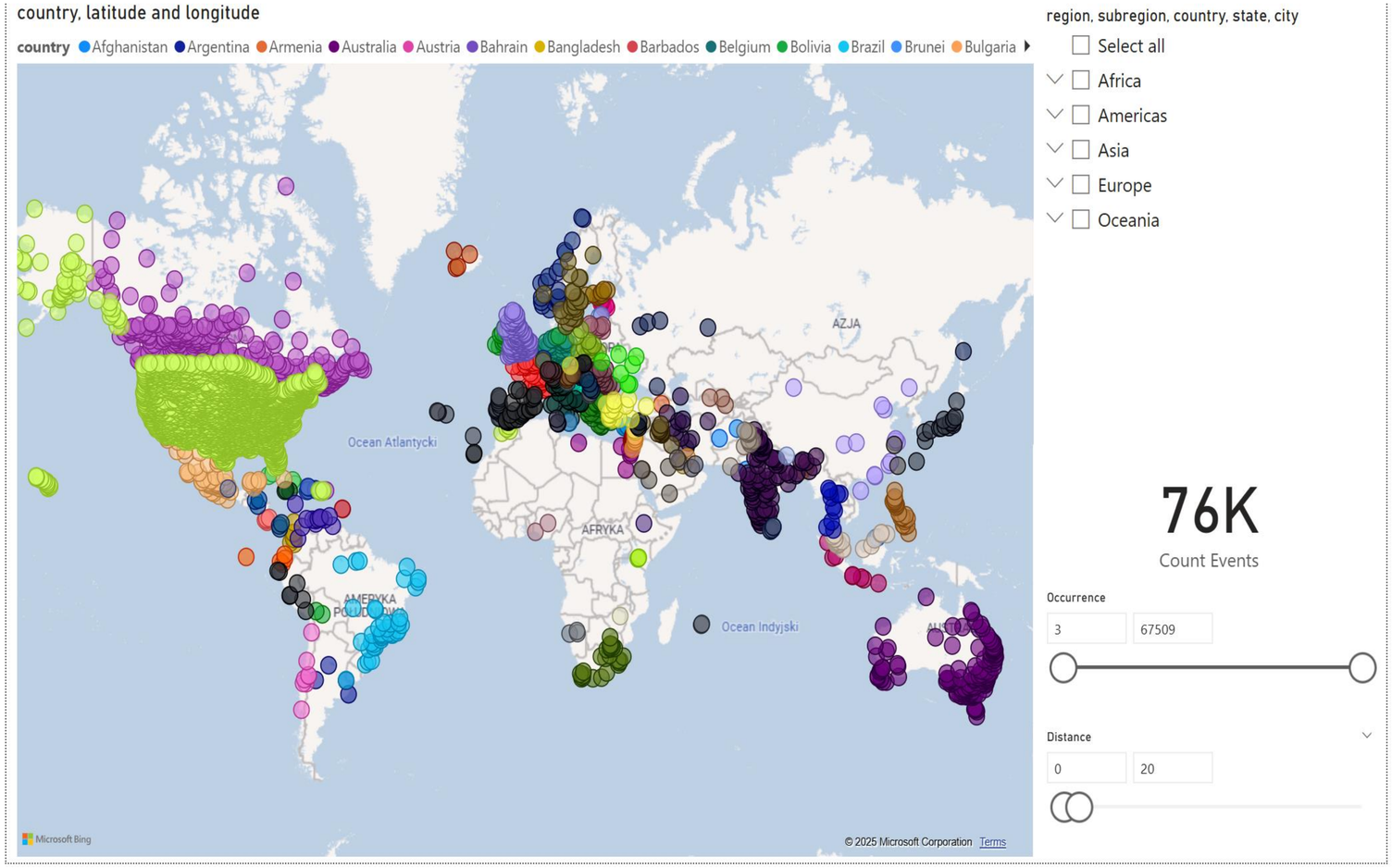
Diagram struktury bazy danych i relacji pomiędzy tabelami



Analytical Data Model

- Analityczny model danych zbudowany w oparciu o tzw. “Star Schema” z centralną tabelą faktów oraz opisowymi tabelami wymiarów.
- Model został nieznacznie rozbudowany celem przeprowadzenia analizy. Dodano min. drugą tabelę faktów dla lokalizacji lotnisk, obie tabele faktów połączono wspólnym wymiarem (“Dim_Country” - Kraj / Lokalizacja geograficzna).
- Model danych wykonany został w aplikacji Power BI Desktop, włączając w to część niezbędnych transformacji danych źródłowych z wykorzystaniem dodatku Power Query.
- Dane źródłowe zostały pobrane z bazy danych SQL, zlokalizowanej na platformie Azure.

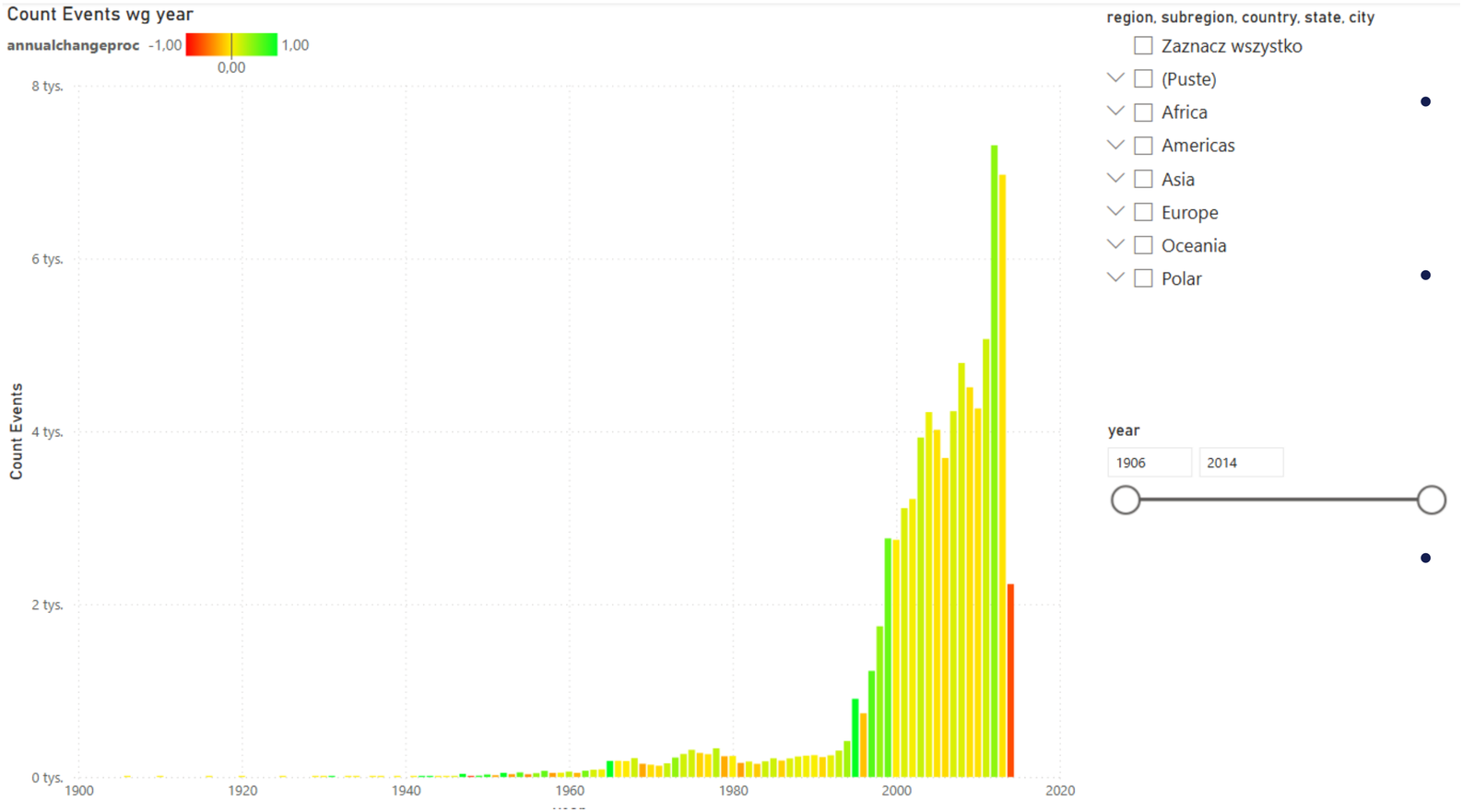
Analiza geograficzna obserwacji UFO



Task_Geographical_Analysis_1

- W analizie uwzględniono 76 tys. zidentyfikowanych od początku XX wieku obiektów UFO
- Najwięcej takich zjawisk zaobserwowano w Ameryce Północnej i Europie oraz południowej Azji
- Najmniej obiektów UFO zidentyfikowano na terenie północnej Azji i Afryce
- Zaobserwowano korelacje między liczbą obserwacji a m.in. gęstością zaludnienia i poziomem rozwoju kraju

Trendy czasowe w obserwacjach UFO (1/3)

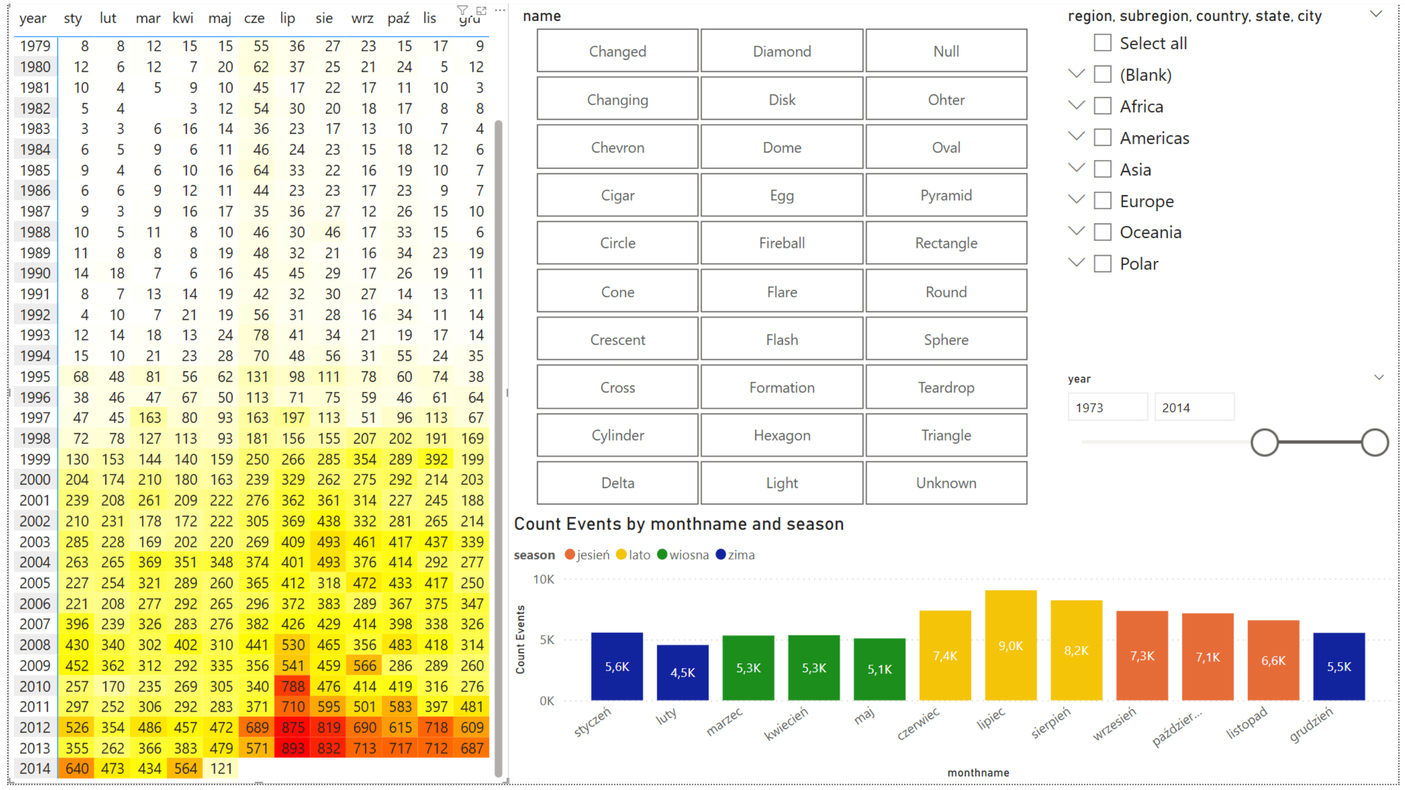


Task_Time_Trend_and_Seasonality_Analysis_1

- W XX wieku nie zaobserwowano wielu zdarzeń związanych z UFO
- Dopiero o drugiej połowy lat 90-tych nastąpił znaczący wzrost zarejestrowanych zjawisk związanych z pojawianiem się UFO
- Największa liczba zjawisk przypada na 2015 rok, kiedy zaobserwowano ponad 7 tysięcy zjawisk, po czym nastąpił znaczący spadek zidentyfikowanych zjawisk

Trendy czasowe w obserwacjach UFO (2/3)

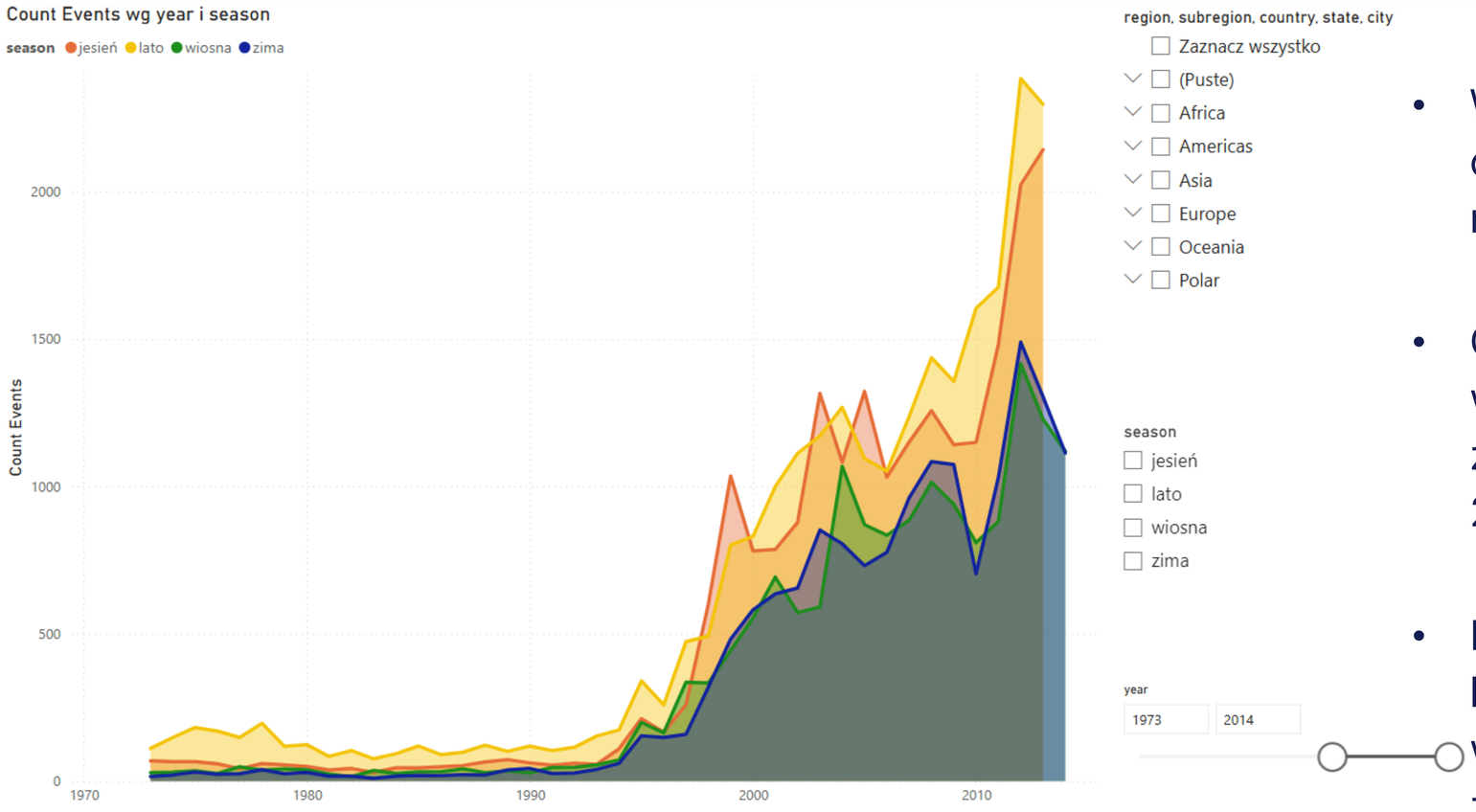
Task_Time_Trend_and_Seasonality_Analysis_2



year	sty	lut	mar	kwi	maj	cze	lip	sie	wrz	paź	lis	gru
1979	8	8	12	15	15	55	36	27	23	15	17	9
1980	12	6	12	7	20	62	37	25	21	24	5	12
1981	10	4	5	9	10	45	17	22	17	11	10	3
1982	5	4		3	12	54	30	20	18	17	8	8
1983	3	3	6	16	14	36	23	17	13	10	7	4
1984	6	5	9	6	11	46	24	23	15	18	12	6
1985	9	4	6	10	16	64	33	22	16	19	10	7
1986	6	6	9	12	11	44	23	23	17	23	9	7
1987	9	3	9	16	17	35	36	27	12	26	15	10
1988	10	5	11	8	10	46	30	46	17	33	15	6
1989	11	8	8	8	19	48	32	21	16	34	23	19
1990	14	18	7	6	16	45	45	29	17	26	19	11
1991	8	7	13	14	19	42	32	30	27	14	13	11
1992	4	10	7	21	19	56	31	28	16	34	11	14
1993	12	14	18	13	24	78	41	34	21	19	17	14
1994	15	10	21	23	28	70	48	56	31	55	24	35
1995	68	48	81	56	62	131	98	111	78	60	74	38
1996	38	46	47	67	50	113	71	75	59	46	61	64
1997	47	45	163	80	93	163	197	113	51	96	113	67
1998	72	78	127	113	93	181	156	155	207	202	191	169
1999	130	153	144	140	159	250	266	285	354	289	392	199
2000	204	174	210	180	163	239	329	262	275	292	214	203
2001	239	208	261	209	222	276	362	361	314	227	245	188
2002	210	231	178	172	222	305	369	438	332	281	265	214
2003	285	228	169	202	220	269	409	493	461	417	437	339
2004	263	265	369	351	348	374	401	493	376	414	292	277
2005	227	254	321	289	260	365	412	318	472	433	417	250
2006	221	208	277	292	265	296	372	383	289	367	375	347
2007	396	239	326	283	276	382	426	429	414	398	338	326
2008	430	340	302	402	310	441	530	465	356	483	418	314
2009	452	362	312	292	335	356	541	459	566	286	289	260
2010	257	170	235	269	305	340	788	476	414	419	316	276
2011	297	252	306	292	283	371	710	595	501	583	397	481
2012	526	354	486	457	472	689	875	819	690	615	718	609
2013	355	262	366	383	479	571	893	832	713	717	712	687
2014	640	473	434	564	121							

- Najwięcej zjawisk związanych z UFO zaobserwowano w miesiącach wakacyjnych 2012 i 2013 roku.
- Najmniej obiektów UFO zarejestrowano w miesiącach styczeń – maj
- Lipiec 2013 był miesiącem gdzie zaobserwowano najwięcej zjawisk, bo aż 893, natomiast w marcu 1982 roku nie odnotowano żadnego pojawienia się obiektu UFO

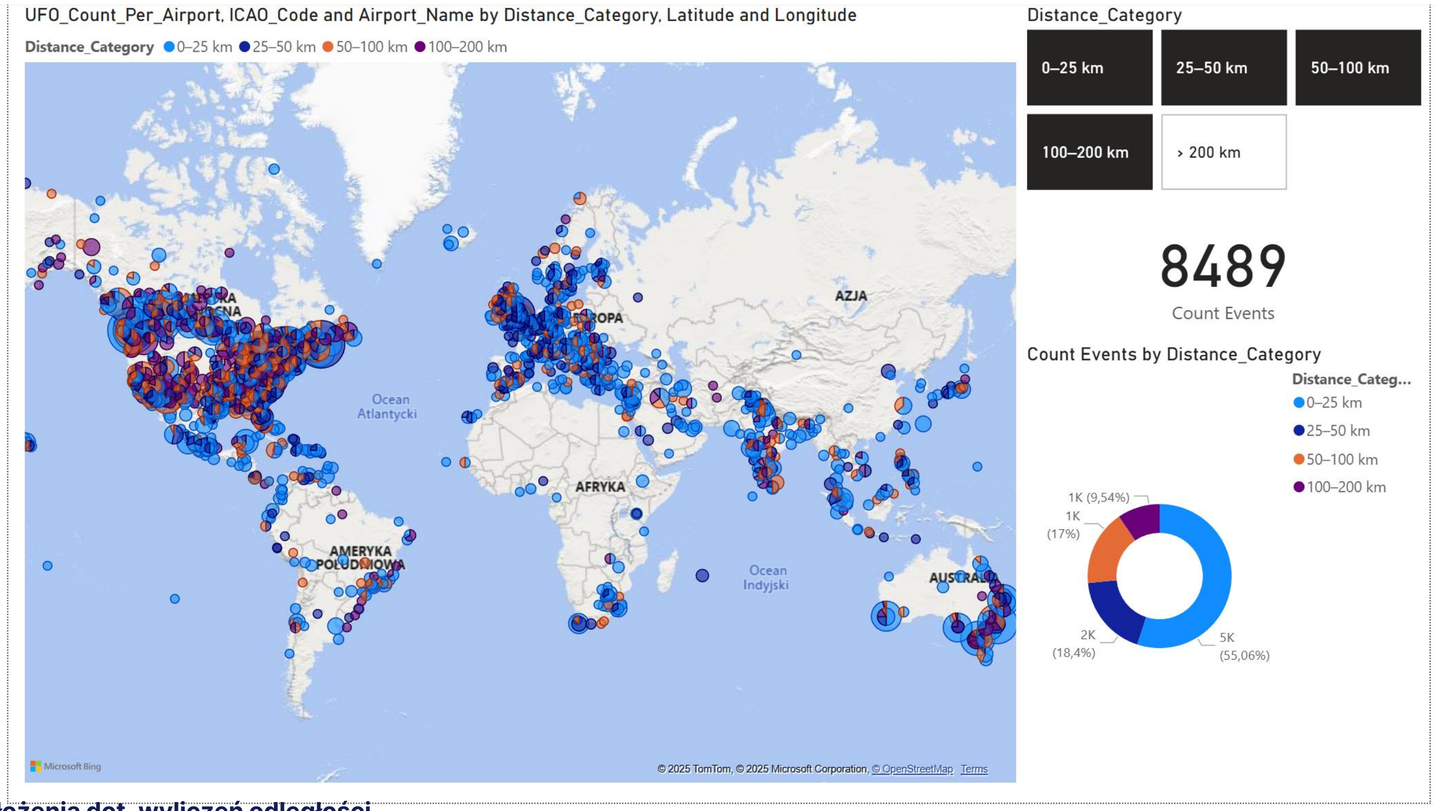
Trendy czasowe w obserwacjach UFO (2/3)



Task_Time_Trend_and_Seasonality_Analysis_3

- W analizie uwzględniono liczbę obserwacji UFO od 1973 do 2014 roku, z podziałem na pory roku
- Od lat 90. obserwuje się wyraźny wzrost liczby raportowanych zjawisk, ze szczytem w okolicach 2012 roku
- Najwięcej zgłoszeń dotyczy okresu letniego, co może wynikać z większej aktywności ludzi na zewnątrz i lepszej widoczności nieba

Korelacja z obiektami geograficznymi (1/2)



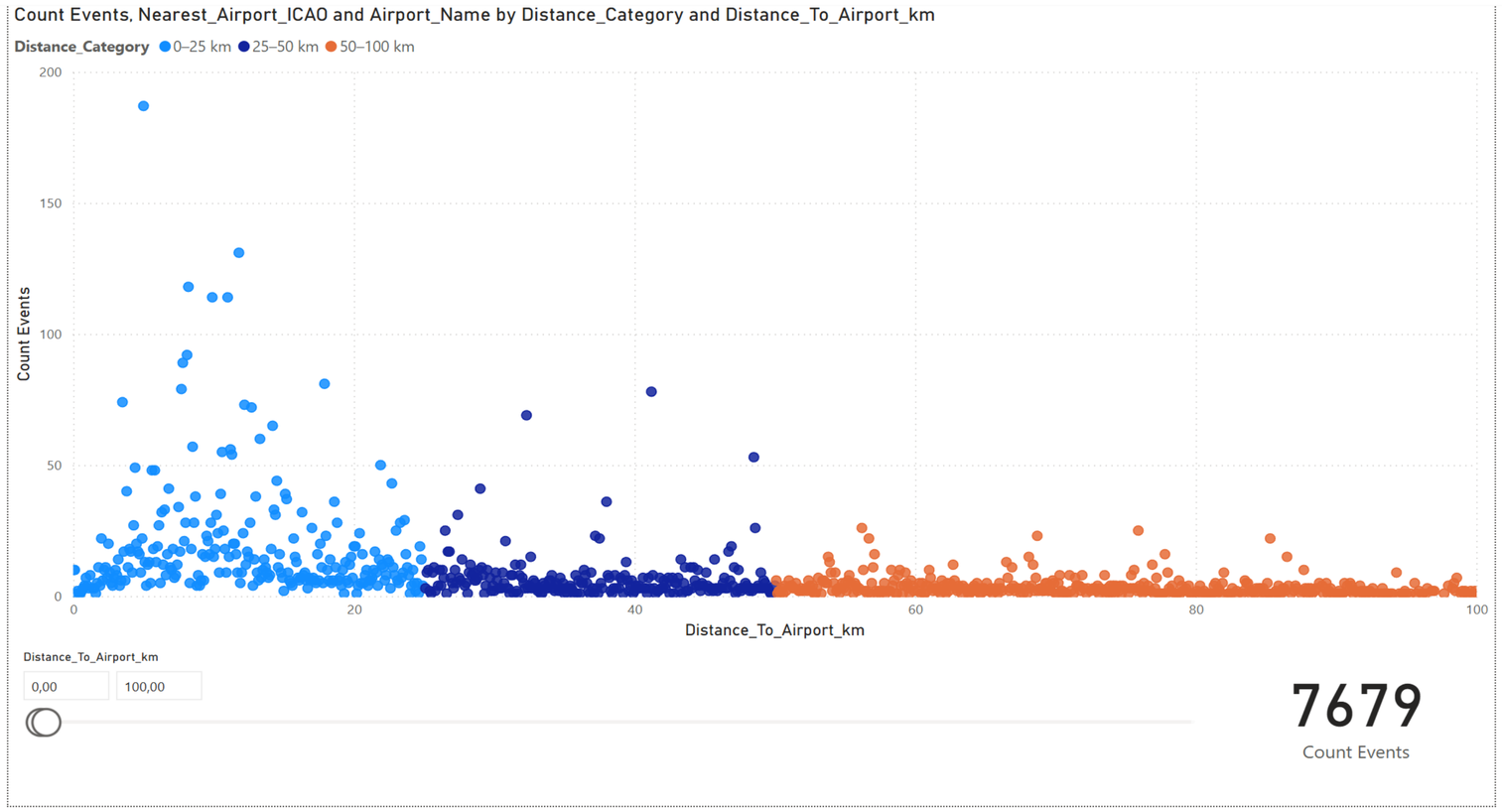
Task_Correlation_Analysis_1

- Większe bańki na mapie oznaczają lokalizacje z dużą liczbą zgłoszeń — najwięcej takich punktów występuje w USA, Europie i Australii, gdzie gęstość infrastruktury lotniczej jest wysoka
- Widoczny jest spadek liczby obserwacji wraz z oddalaniem się od lotnisk: 0-25 km (55,06%), 25–50 km (18,4%), 50–100 km (17%) i 100–200 km (9,5%)

Założenia dot. wyliczeń odległości

- Odległość obliczana jest za pomocą wzoru Euklidesowego w układzie współrzędnych, nie uwzględniając zakrzywienia Ziemi.
- Wynik jest skalowany przez stałą “111.32 km” na 1 stopień (średnia wartość 1° geograficznego na powierzchni Ziemi).
- Wynik końcowy jest zaokrąglany do 1 miejsca po przecinku.
- Zdecydowaliśmy się na określenie odległości w linii prostej z uwagi na względnie małe odległości pomiędzy obliczanymi punktami (korelacja zjawisko i lotnisko do 200km), pomogło nam to również uprościć konieczne obliczenia.

Korelacja z obiektami geograficznymi (2/2)

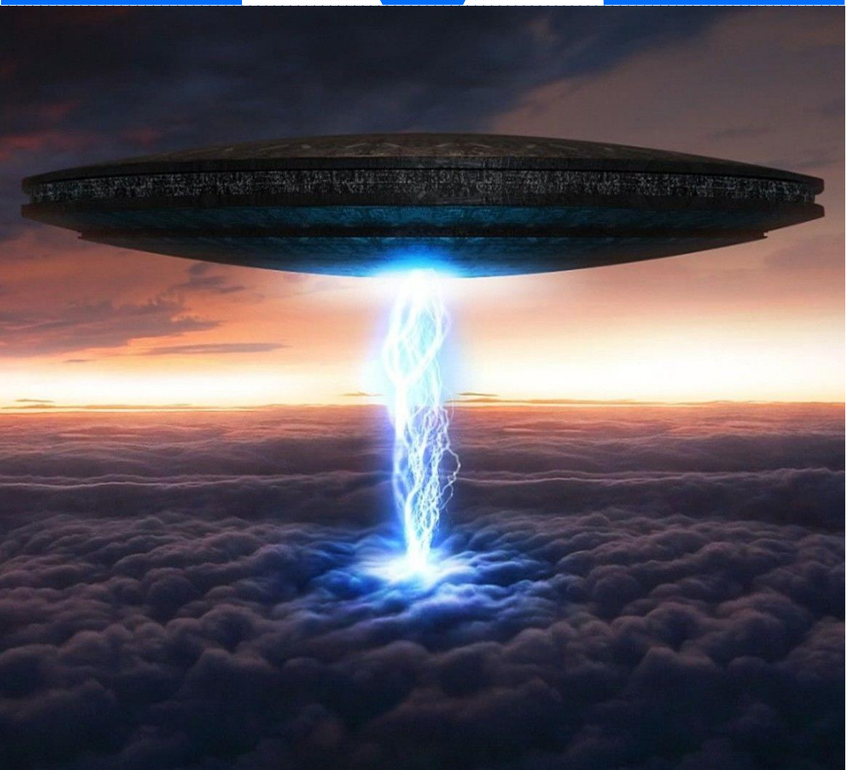


Task_Correlation_Analysis_2

- Uwzględniono 7 679 obserwacji UFO w promieniu do 100 km od lotnisk
- Najwięcej zgłoszeń dotyczy obszarów do 25 km od lotniska, a ich liczba spada wraz z odległością
- Zjawiska te mogą być mylone z samolotami lub helikopterami, co może wpływać na liczbę raportów w pobliżu infrastruktury lotniczej

Linki

- **Azure DevOps Repos:** https://dev.azure.com/ppx34654/Dyplom%20WSB%20-%20UFO
- **Git Hub**: https://github.com/admiralRobson/DyplomUFO



"Dzięki mocy Big Data spróbowaliśmy odpowiedzieć na pytanie, czy kosmici naprawdę wolą lądować w Teksasie, a do tego sprawdzaliśmy, czy UFO pojawia się częściej w wakacje – bo nawet obcy mogą mieć wolne.
No i chyba nam się to udało ;)."