1_data_verkennen_final

December 11, 2025

1 Data verkennen

[17]:
```
from pyspark.sql import SparkSession
```

[3]:
```
spark = SparkSession.builder \
    .appName("Drone Project Exploratie") \
    .getOrCreate()

# Stap 2: Lees de CSV
# header=True: De eerste rij bevat kolomnamen
# inferSchema=True: Spark raadt zelf of iets tekst of een getal is
df = spark.read.csv("/home/jovyan/work/drone-project/cnas_drone_proliferation.
 ↪csv", header=True, inferSchema=True)

# Stap 3: Bekijk de structuur (schema)
# Dit vertelt je welke kolommen je hebt en welk type (String, Integer) ze zijn
df.printSchema()

# Stap 4: Bekijk de eerste 5 rijen netjes
df.show(5, truncate=False)
```

```
root
 |-- Entry Number : string (nullable = true)
 |-- Region : string (nullable = true)
 |-- Seeker: string (nullable = true)
 |-- Supplier: string (nullable = true)
 |-- Paltform Model: string (nullable = true)
 |-- Platform  Type: string (nullable = true)
 |-- Year of Interest : double (nullable = true)
 |-- Year of Order: double (nullable = true)
 |-- Year of First Delivery: double (nullable = true)
 |-- Year of  Identification, Completion, or Cancellation: double (nullable =
true)
 |-- Status: string (nullable = true)
 |-- Source 1: string (nullable = true)
 |-- Source 2: string (nullable = true)
 |-- Source 3: string (nullable = true)
 |-- Comments: string (nullable = true)
```

```
+------------+-------------+-------------+--------+-------------+-------------
-+----------------+-------------+---------------------+----------------------
----------------------------+------+--------+--------+--------+---------------
-------------------------------------------------------------------------------
-------------------------------------------------+
|Entry Number |Region       |Seeker       |Supplier|Paltform Model|Platform
Type|Year of Interest |Year of Order|Year of First Delivery|Year of
Identification, Completion, or Cancellation|Status|Source 1|Source 2|Source
3|Comments
|
+------------+-------------+-------------+--------+-------------+-------------
-+----------------+-------------+---------------------+----------------------
----------------------------+------+--------+--------+--------+---------------
-------------------------------------------------------------------------------
-------------------------------------------------+
|1            |North America|United States|Israel  |Hunter        |UAV
|NULL             |1992.0       |1993.0               |1996.0
|Yes   |SIPRI   |NULL    |NULL    |"SIPRI Comments: ""Deal incl. 7 ground-
control systems; ordered via U.S. company; possibly assembled in the United
States; U.S. designation RQ-5A"""|
|2            |Indo-Pacific |Singapore    |Israel  |Searcher Mk II|UAV
|NULL             |1993.0       |1994.0               |1995.0
|Yes   |SIPRI   |NULL    |NULL    |NULL
|
|3            |Europe       |France       |Israel  |Hunter        |UAV
|NULL             |1995.0       |1997.0               |1997.0
|Yes   |SIPRI   |NULL    |NULL    |"SIPRI Comments: ""$50m deal"""
|
|4            |Europe       |Netherlands  |France  |Sperwer       |UAV
|NULL             |1995.0       |2000.0               |2001.0
|Yes   |SIPRI   |NULL    |NULL    |"SIPRI Comments: ""$82m deal; retired in
2011"""
|
|5            |Indo-Pacific |Sri Lanka    |Israel  |Scout         |UAV
|NULL             |1995.0       |1998.0               |1998.0
|Yes   |SIPRI   |NULL    |NULL    |"SIPRI Comments: ""Secondhand but modernized
before delivery; for use against LTTE rebels; $4.3m deal; Super Scout version"""
|
+------------+-------------+-------------+--------+-------------+-------------
-+----------------+-------------+---------------------+----------------------
----------------------------+------+--------+--------+--------+---------------
-------------------------------------------------------------------------------
-------------------------------------------------+
only showing top 5 rows
```

1.1 Conclusie:

1.1.1 Belangrijke features:

- Supplier vs. Seeker: Dit is de kern van je netwerk-analyse.

Doel: We kunnen hiermee de "flow" van wapens in kaart brengen (wie bevoorraadt wie?).

- Platform Type:

Waarde: Hiermee maken we onderscheid tussen simpele verkenners (UAV) en dodelijke wapens (Armed UAV / Loitering Munition).

Trend: De verschuiving van UAV naar Loitering Munition is precies wat je "Evolution" deel moet aantonen.

- Year of Order / Year of First Delivery:

Waarde: Dit is je tijdsas. Zonder dit kunnen we geen evolutie tonen.

1.1.2 Eerst wat technische fixes:

- Er staat Paltform Model in plaats van Platform Model
- De jaren (bv. Year of Order) zijn double (bv. 1992.0). Voor een visualisatie willen we liever een geheel getal (Integer: 1992) of een datum
- nullable = true betekent dat data kan ontbreken. Bij visualisaties moeten we straks beslissen: laten we lege jaren weg of vullen we ze op?

[4]:
```
from pyspark.sql.functions import col

# 1. Kolom hernoemen (Fix de typo)
# We maken een nieuwe DataFrame 'df_clean' zodat we de originele 'df' niet␣
 ↪kwijt zijn
df_clean = df.withColumnRenamed("Paltform Model", "Platform Model")

# 2. Jaren omzetten van Double (1995.0) naar Integer (1995)
# We doen dit alleen voor de kolommen die we echt gaan gebruiken
df_clean = df_clean.withColumn("Year of Order", col("Year of Order").
 ↪cast("integer")) \
                   .withColumn("Year of First Delivery", col("Year of First␣
 ↪Delivery").cast("integer"))

# 3. Eerste Analyse: Wie zijn de grootste leveranciers?
print("Top 5 Drone Leveranciers (Suppliers):")
df_clean.groupBy("Supplier") \
    .count() \
    .orderBy(col("count").desc()) \
    .show(5)

# 4. Tweede Analyse: Wat voor type drones worden verhandeld?
print("Verdeling per Type:")
df_clean.groupBy("Platform  Type") \
```

```
    .count() \
    .show(truncate=False)
```

```
Top 5 Drone Leveranciers (Suppliers):
+-------------+-----+
|     Supplier|count|
+-------------+-----+
|United States|  198|
|       Israel|  190|
|       Turkey|   69|
|         Iran|   56|
|        China|   55|
+-------------+-----+
only showing top 5 rows
Verdeling per Type:
+------------------+-----+
|Platform  Type    |count|
+------------------+-----+
|Armed UAV         |146  |
|NULL              |12   |
|UAV               |513  |
|Armed UAV         |1    |
|Loitering Munition|60   |
+------------------+-----+
```

1.2 Conclusie 2

1.2.1 Wie zijn de grootste leveranciers ?

1. United States
2. Israel
3. Turkije
4. Iran
5. China

1.2.2 Welke types drones worden vooral verhandeld ?

- Loitering Munition –> Dit zijn kamikaze drones.

[ ]:

2 Visualisaties

[5]:
```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
from pyspark.sql.functions import col, count, when
```

```
# Zet de styling van seaborn vast
sns.set_theme(style="whitegrid")
```

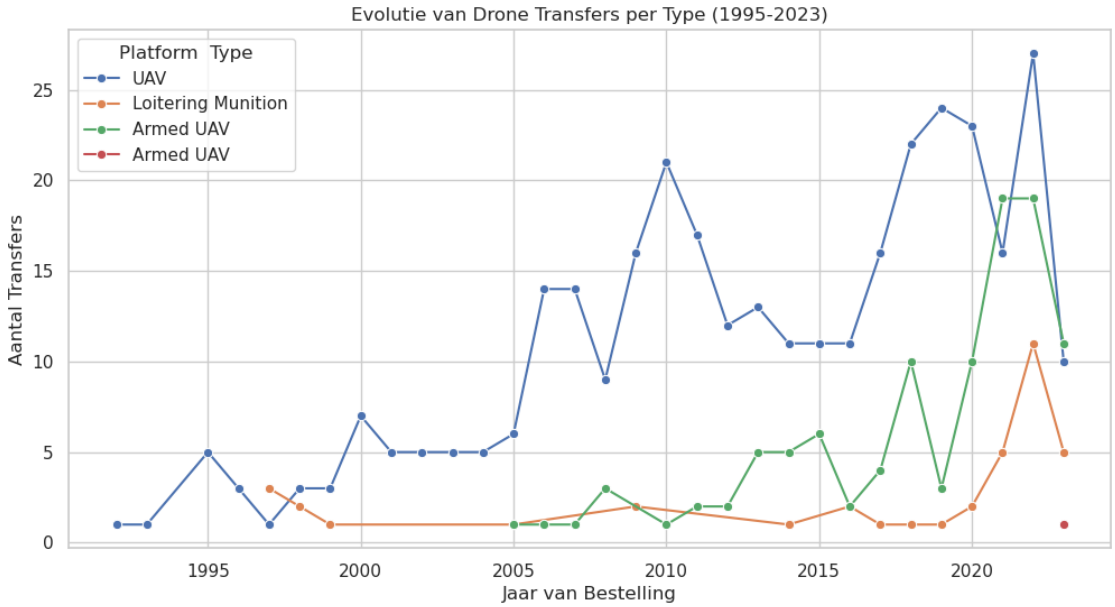
2.0.1 Kamikaze Drones (Loitering Munition) in de tijd

[6]:
```
# --- STAP 1: Data Aggregatie in Spark ---
# We groeperen per Jaar en Type, en tellen het aantal entries
df_trend = df_clean.filter(col("Year of Order").isNotNull()) \
    .groupBy("Year of Order", "Platform  Type") \
    .count() \
    .orderBy("Year of Order")

# Zet om naar Pandas voor visualisatie
pdf_trend = df_trend.toPandas()

# --- STAP 2: Visualisatie ---
plt.figure(figsize=(12, 6))
sns.lineplot(data=pdf_trend, x="Year of Order", y="count", hue="Platform ␣
 ↪Type", marker="o")

plt.title("Evolutie van Drone Transfers per Type (1995-2023)")
plt.ylabel("Aantal Transfers")
plt.xlabel("Jaar van Bestelling")
plt.show()
```



Conclusie: We zien dat er een sterke stijging is in de orders en leveringen van Loitering Munition vanaf 2020. Dit suggereert een groeiende interesse en vraag naar deze specifieke drone-technologie in de afgelopen jaren. De piek in 2022 kan mogelijk worden toegeschreven aan geopolitieke spanningen (oorlog oekraine) of conflicten die de vraag naar dergelijke wapensystemen hebben aangewakkerd.

2.0.2 Evolutie drone transfers per land

We zullen proberen waarnemen of er buiten VS en China een opkomst is van nieuwe leveranciers van drones.

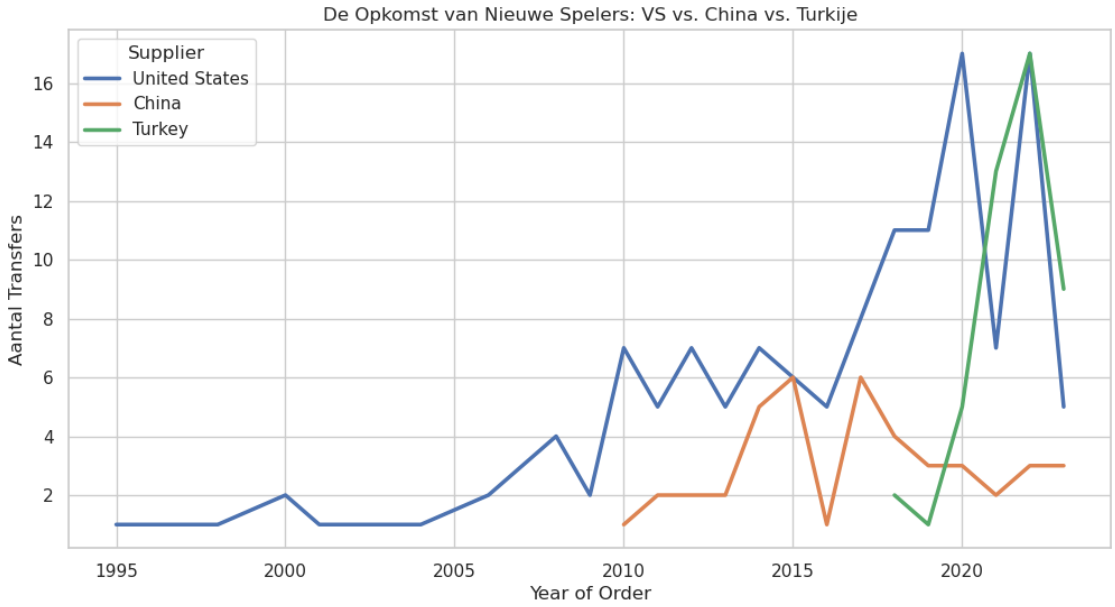
[7]:
```
# --- STAP 1: Filteren op Top 3 Suppliers ---
top_suppliers = ["United States", "China", "Turkey"]

# Filter de dataset in Spark
df_suppliers = df_clean.filter(col("Supplier").isin(top_suppliers)) \
    .filter(col("Year of Order").isNotNull()) \
    .groupBy("Year of Order", "Supplier") \
    .count() \
    .orderBy("Year of Order")

# Naar Pandas
pdf_suppliers = df_suppliers.toPandas()

# --- STAP 2: Visualisatie ---
plt.figure(figsize=(12, 6))
sns.lineplot(data=pdf_suppliers, x="Year of Order", y="count", hue="Supplier",␣
 ↪linewidth=2.5)

plt.title("De Opkomst van Nieuwe Spelers: VS vs. China vs. Turkije")
plt.ylabel("Aantal Transfers")
plt.show()
```



Conclusie:

- We zien dat de VS en China dominant zijn in de drone-markt, maar er is een duidelijke opkomst van andere landen zoals Turkije en Iran als leveranciers van drones. Deze landen hebben hun productiecapaciteit en technologische expertise vergroot, wat hen in staat stelt om een grotere rol te spelen in de wereldwijde drone-markt. We weten dat Turkije en Iran actief zijn in conflicten in het Midden-Oosten, wat hun behoefte aan geavanceerde drone-technologie kan verklaren.

2.0.3 Marktaandeel per jaar

(2005-2023) van de top 5 leveranciers We gebruiken hiervoor een barchart omdat deze visualisatie het beste de veranderingen in proportie van drone transfers per jaar kan weergeven. Lijngrafieken zijn goed voor trends. We verwachten hier dus dat Turkije en Iran een groter aandeel krijgen in de loop der jaren.

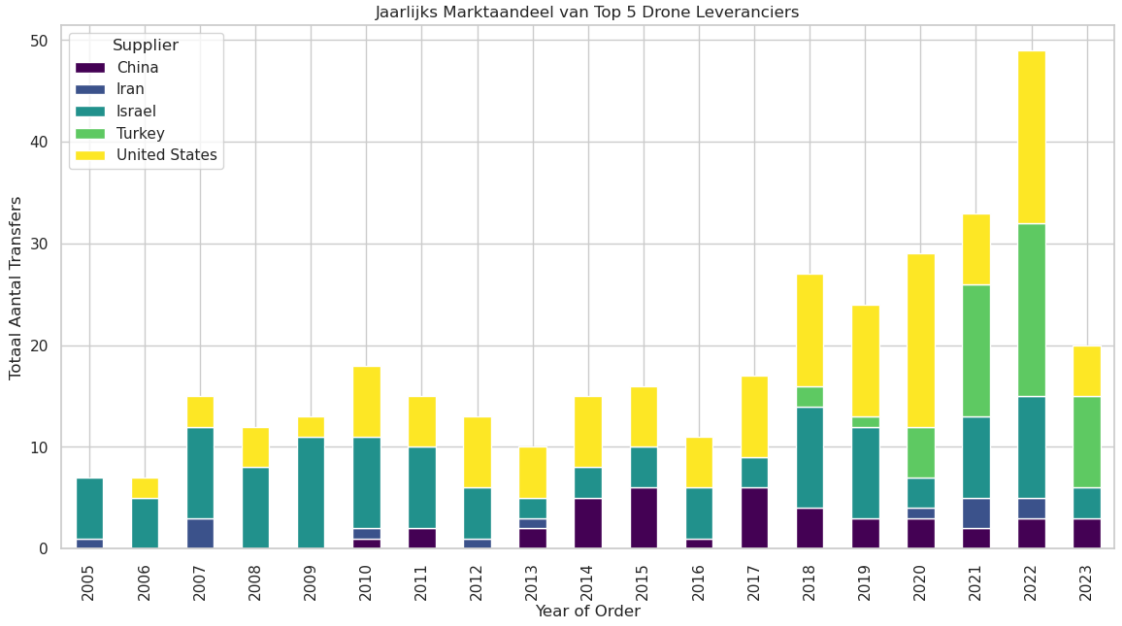
[13]:
```
# --- STAP 1: Data voorbereiden (Top 5 suppliers voor leesbaarheid) ---
# We nemen de top 5 suppliers, anders wordt de grafiek te druk
top5_list = [row['Supplier'] for row in df_clean.groupBy("Supplier").count().
 ↪orderBy(col("count").desc()).limit(5).collect()]

df_market = df_clean.filter(col("Supplier").isin(top5_list)) \
    .filter(col("Year of Order") >= 2005) \
    .groupBy("Year of Order", "Supplier") \
    .count() \
    .orderBy("Year of Order")
```

```
# Pivot table in Pandas maken voor stacked bar
pdf_market = df_market.toPandas()
pdf_pivot = pdf_market.pivot(index='Year of Order', columns='Supplier',␣
 ↪values='count').fillna(0)

# --- STAP 2: Visualisatie ---
pdf_pivot.plot(kind='bar', stacked=True, figsize=(14, 7), colormap='viridis')

plt.title("Jaarlijks Marktaandeel van Top 5 Drone Leveranciers")
plt.ylabel("Totaal Aantal Transfers")
plt.show()
```



Conclusie: We kunnen hier letterlijk zien hoe het gekleurde blok van Turkije (Groen) ineens omhoog schiet rond 2021-2022. Dit bevestigt onze hypothese dat Turkije een opkomende speler is in de drone-markt. Iran (Paars) laat ook een lichte stijging zien, wat aangeeft dat zij ook hun aanwezigheid in deze markt uitbreiden. We stellen ook vast dat van 2000 tot 2004 er eigenlijk maar 2 spelers in de drone markt waren: Israel en VS. Dit toont aan hoe geconcentreerd de markt vroeger was, in tegenstelling tot de meer diverse markt van vandaag.

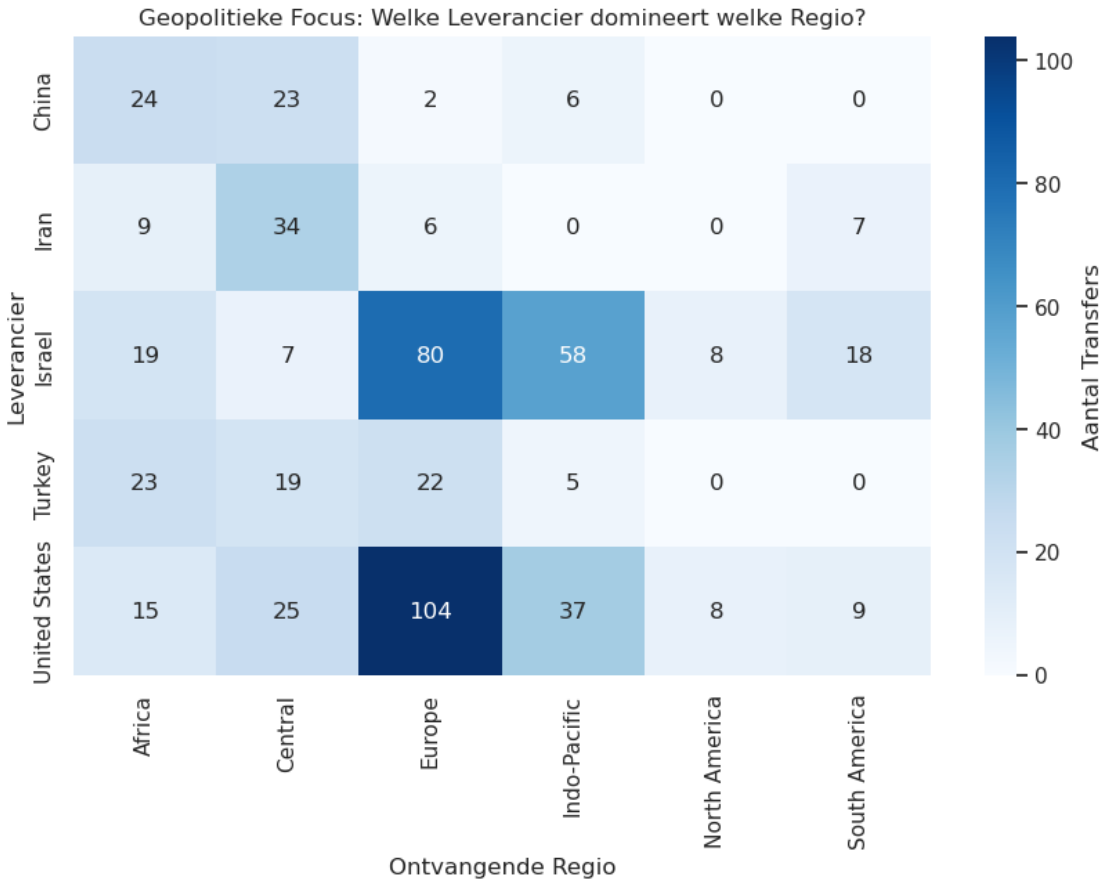
2.0.4 Wie levert aan wie ?

[16]:
```
# --- STAP 1: Cross-tabulatie in Spark ---
# We groeperen op Supplier EN Region
df_geo = df_clean.filter(col("Supplier").isin(top5_list)) \
    .groupBy("Supplier", "Region ") \
    .count()
```

```
# Naar Pandas en Pivotten (Matrix vorm maken)
pdf_geo = df_geo.toPandas()
matrix_geo = pdf_geo.pivot(index="Supplier", columns="Region ", values="count").
 ↪fillna(0)

# --- STAP 2: Visualisatie ---
plt.figure(figsize=(10, 6))
sns.heatmap(matrix_geo, annot=True, fmt='g', cmap="Blues", cbar_kws={'label':␣
 ↪'Aantal Transfers'})

plt.title("Geopolitieke Focus: Welke Leverancier domineert welke Regio?")
plt.xlabel("Ontvangende Regio")
plt.ylabel("Leverancier")
plt.show()
```



Conclusie We kunnen hier duidelijk zien dat : - VS veel levert aan “Europa” (NAVO landen) en aan Indo Pacific (waarschijlijk landen zoals Thailand). - China levert vooral aan Afrika en Azië. - Iran levert vooral aan Midden-Oosten. (alhoewel we bij Iran wel veel “Unknown” zien staan als

seeker, dus dit kan vertekenen) Zoals het code book al zegt: “Nonstate actors and authoritarian regimes have significant motivation to keep details… secret” - Turkije levert aan Europa en Midden Oosten.