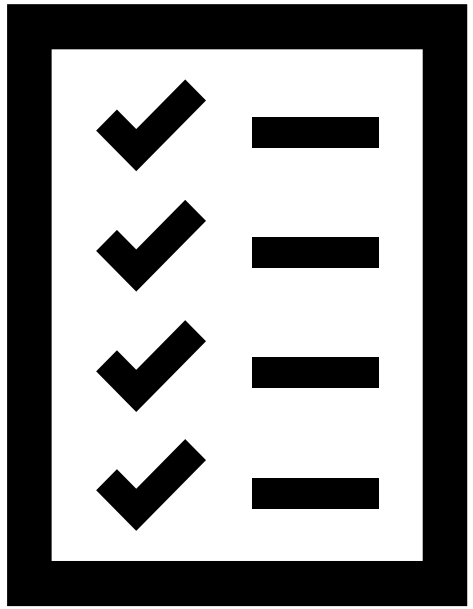
INDUSTRY SPONSOR

Country Web Intelligence Mining

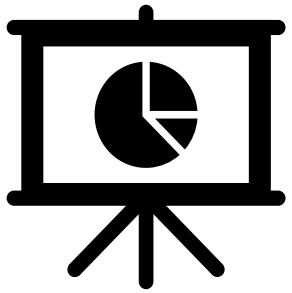
PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

OVERVIEW

1. Business Problem
2. Assumptions
3. Approach + Decisions Taken
4. Code Operations + Outputs
5. Processing and Automation
6. PowerBI Dashboard Outputs Demo



BUSINESS PROBLEM



Problem Breakdown:

- **Business Need:** Automate country intelligence gathering to identify key defense activities.
- **Analytics Approach:** Leverage web scraping, dashboards, and machine learning to extract and rank relevant defense insights.

Potential Solutions?

- **Web Scraping:** *BeautifulSoup, Selenium,* or *Microsoft's Power Automate*
- **Visualization:** *PowerBI* or *Tableau* for dashboards that can be converted to PPT.
- **ML Approach:** Sentiment analysis for more in depth insights

REQUIREMENTS GATHERING

BIG QUESTIONS WE ASKED OURSELVES?

- **Where** will our data come from?
  - What websites?
- **How** will we structure the data?
  - Coding in specific formats for easy-to-read file outputs? Can PowerBI or Tableau understand the outputs?
- **Can** we script the data to be pulled through one easy push of a button?
  - Automation in the code processes? Making it as easy as possible to use? Army-proof?
- **What** will our data look like in a visualization tool?
  - Is it easy to use? What insights can be extracted? Are they useful?

PICKING A WAY FORWARDS

- We had two paths to choose from at an early stage:

1	2
Utilize and improve upon Python library scraping methods through either BeautifulSoup or Selenium packages	Create a workflow process through an agentic AI process (autonomous task completing intelligence)
PROs: Proven time and time again, easy to use and learn, very VERY flexible	PROs: Straightforwards and easy to build a flow – no real coding experience needed
CONs: Requires some coding understanding to properly use, can be complex depending on websites being scraped	CONs: Microsoft Power Automate poses challenges like a learning curve, issues with GPT scraping defense data, and unreliable data since it's sourced via GPT.

ATTEMPTING AGENTIC AI

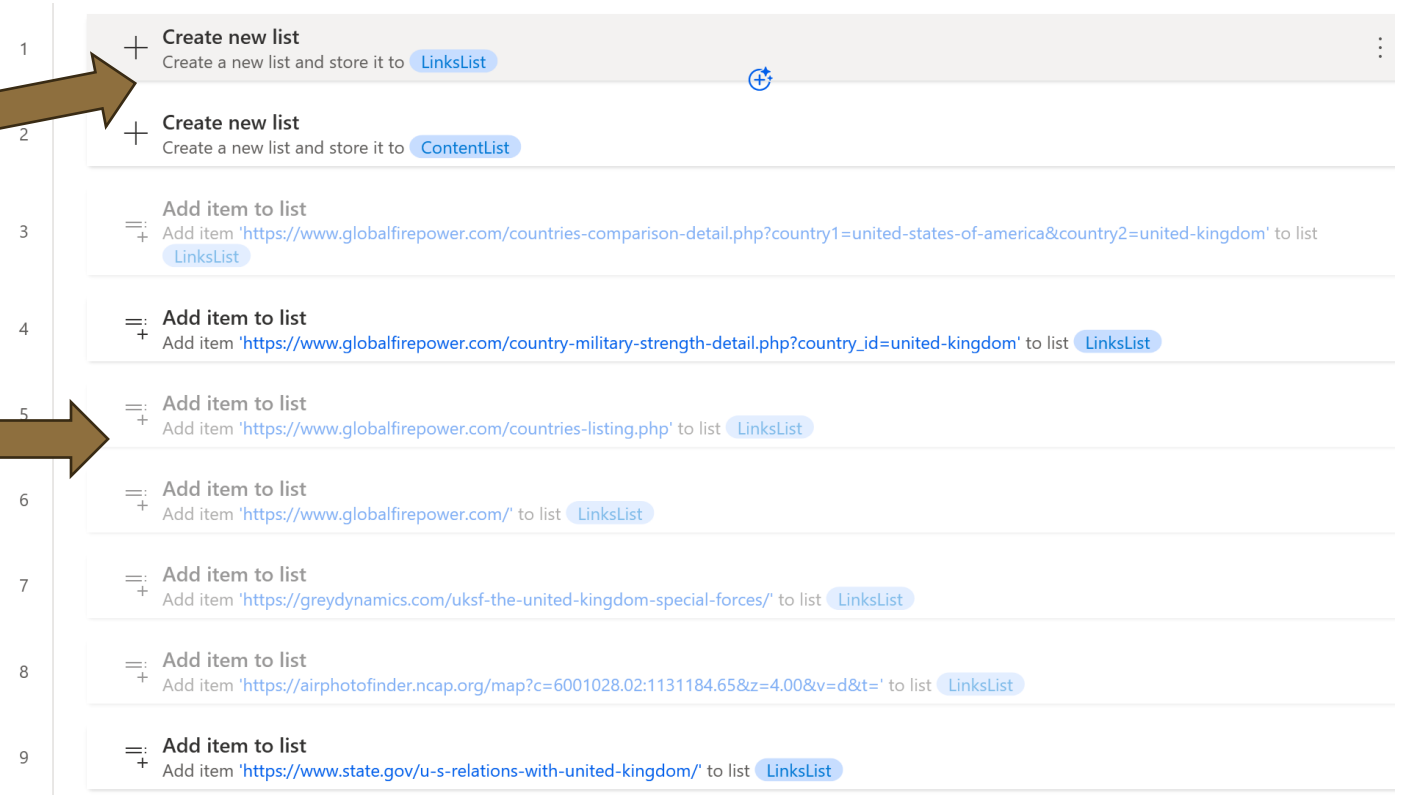
MICROSOFT POWER AUTOMATE

Step 1

Create a list

Step 2

Add website URLs to list



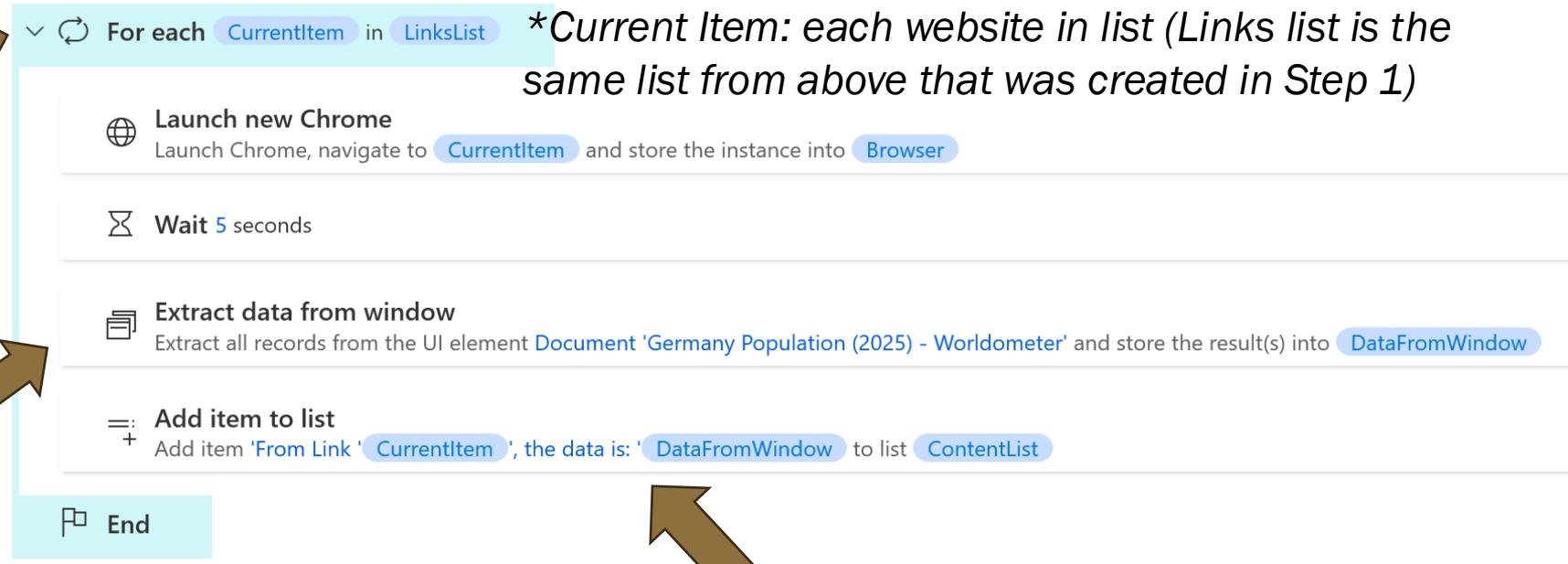
MICROSOFT POWER AUTOMATE

Step 3

Create for each loop (same concept as Python For loop), Launch Chrome for scraping.

Step 4

Extract data from window function and store it into a variable (DataFromWindow)



Current Item: each website in list (Links list is the same list from above that was created in Step 1)

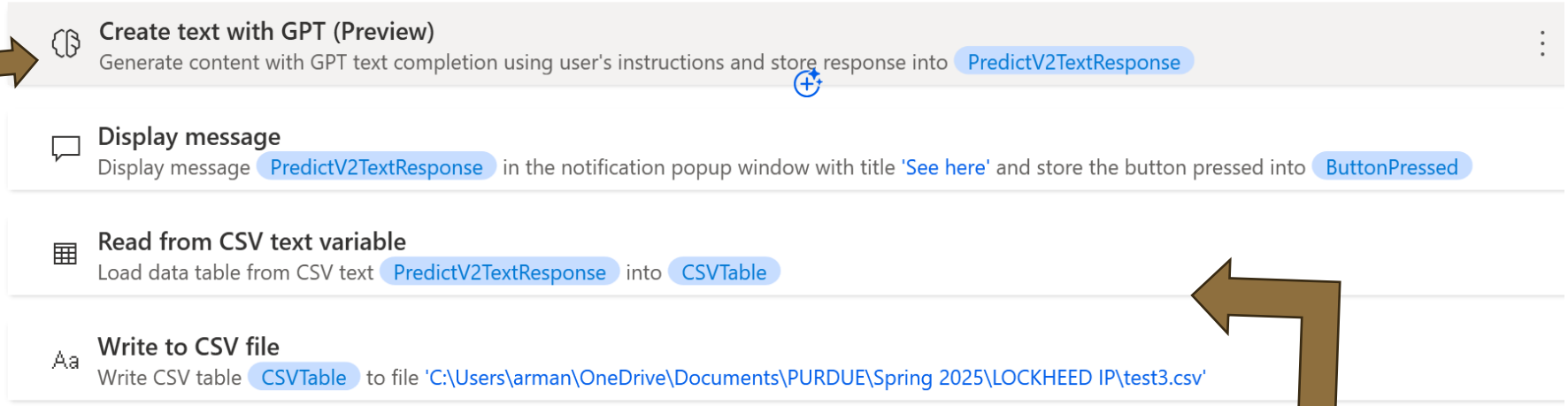
Step 5

Add function and store it into a variable (DataFromWindow). Associate each website link in the Linkslist to an extracted data pull.

MICROSOFT POWER AUTOMATE

Step 6

Utilizing in-built AI GPT to prompt CSV creation



Instructions:

This data is all about the United Kingdom. Combine the data from this list into one compete CSV ready for exporting. If the data doesnt make sense and would not be helpful to help a team trying to sell a product in the United Kingdom (UK), don't include it. Just the CSV data, nothing else. Please make it easy to read and ensure units are spelled out. List: %ContentList%

Step 6

Load and write the file into CSV format.

ISSUES WITH MICROSOFT POWER AUTOMATE

- Microsoft's GPT *not flexible* with inputs – lots of issues running the flows
- Power Automate's user interface is *very specific* and requires training to understand fully.
- Time savings are *not noticeable* enough to warrant a full switch over

OUR CONCLUSION -- The **MAIN** issue is the **GPT**, it cannot properly process inputs regarding defense data being extracted for CSV outputs.

PYTHON WEB SCRAPING

WEB SCRAPING WITH BEAUTIFULSOUP + SELENIUM

- **FIRST,** we split up to focus on specific sections of data to scrape for each country:

Such as -- Defense/Economics/Demographics/and others…

- **SECOND,** we each created workflow processes and assumptions for our respective sections:

To help reproduce our results

- **THIRD,** we combined our codes and dashboards into a centralized location for use:

Easy code script to run from one place/centralized github repository/one dashboard on PowerBI

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

SOURCES – WEBSITES WE SCRAPED DATA FROM

For our project, we collected data from various reliable public sources, including:

🌍 **Economic & Trade Data:**

- Trading Economics (https://tradingeconomics.com/countries)
- Macrotrends (https://www.macrotrends.net/global-metrics/countries)

🛡️ **Military Data:**

- Global Firepower (https://www.globalfirepower.com/)
- World Directory of Modern Warships (https://www.wdmmw.org/)

🌿 **Natural Resources & Geography:**

- CIA World Factbook (https://www.cia.gov/the-world-factbook/)
- UN Environment Programme (https://www.unep.org/)

🌦️ **Climate Data:**

- NOAA Climate Data (https://www.ncdc.noaa.gov/)
- World Meteorological Organization (https://public.wmo.int/en)

DATA PROCESSING + AUTOMATION

```
CHROME_DRIVER_PATH = r"C:\Web Driver\chromedriver.exe"
BASE_URL = "https://www.militaryfactory.com/armor/by-country.php?Nation="

country_list = [
    "belarus", "uruguay", "jordan", "botswana", "taiwan", "central-african-republic", "niger", "ethiopia",
    "oman", "algeria", "chad", "south-sudan", "mozambique", "austria", "germany", "nicaragua", "japan",
    "united-arab-emirates", "belgium", "paraguay", "canada", "peru", "libya", "somalia", "north-macedonia",
    "tajikistan", "france", "turkey", "venezuela", "romania", "bulgaria", "angola", "kenya", "thailand",
    "democratic-republic-of-the-congo", "latvia", "saudi-arabia", "denmark", "cuba", "guatemala",
    "el-salvador", "spain", "mali", "suriname", "india", "vietnam", "israel", "georgia", "philippines",
    "slovenia", "chile", "ivory-coast", "sweden", "colombia", "republic-of-the-congo", "qatar", "eritrea",
```

Figuring out website URLs to create efficient scrapers

Creating neat CSV outputs

```
# Save to CSV
df = pd.DataFrame(all_armor_data, columns=["Country", "System Name", "Role", "Year", "Image URL"])
df.to_csv(r"C:\Users\arman\OneDrive\Documents\PURDUE\Spring 2025\LOCKHEED IP\armor_images.csv", index=False)
print("✅ Scraping complete! Data saved to 'armor_inventory_with_thumbnails.csv'.")
```

DATA PROCESSING + AUTOMATION

	A	B	C	D	E
1	Country	Aircraft	Units	Role	
2	Mexico	F-5E	6	Fighter	
3	Mexico	PC-7	33	Light Attack	
4	Mexico	UH-60M	20	Multi-Mission	
5	Mexico	Mi-8/-17	18	Transport/Gunship	
6	Mexico	Bell 407	17	Light Utility	
7	Mexico	H225M	16	SAR/Utility	
8	Mexico	MD530F	13	Light Utility	
9	Mexico	Bell 206	13	Medium Utility	
10	Mexico	Bell 412	8	Medium Utility	
11	Mexico	UH-1H	1	Medium Utility	
12	Mexico	C-295	7	Tactical	
13	Mexico	C-27J	4	Tactical	
14	Mexico	C-130E/K/L	3	Tactical	
15	Mexico	King Air 90/	3	Utility	
16	Mexico	Boeing 737	3	VIP	
17	Mexico	Turbo Com	2	Utility	
18	Mexico	PC-6	1	Utility	
19	Mexico	T-6C+	56	Basic Trainer	
20	Mexico	PC-7	33	Basic/Advanced Trainer	

- **GOAL** create clean and easily digestible CSV files that can be transferred to visualization tools
- **SMOOTH** data transfers to decrease time spent debugging and fixing CSV files

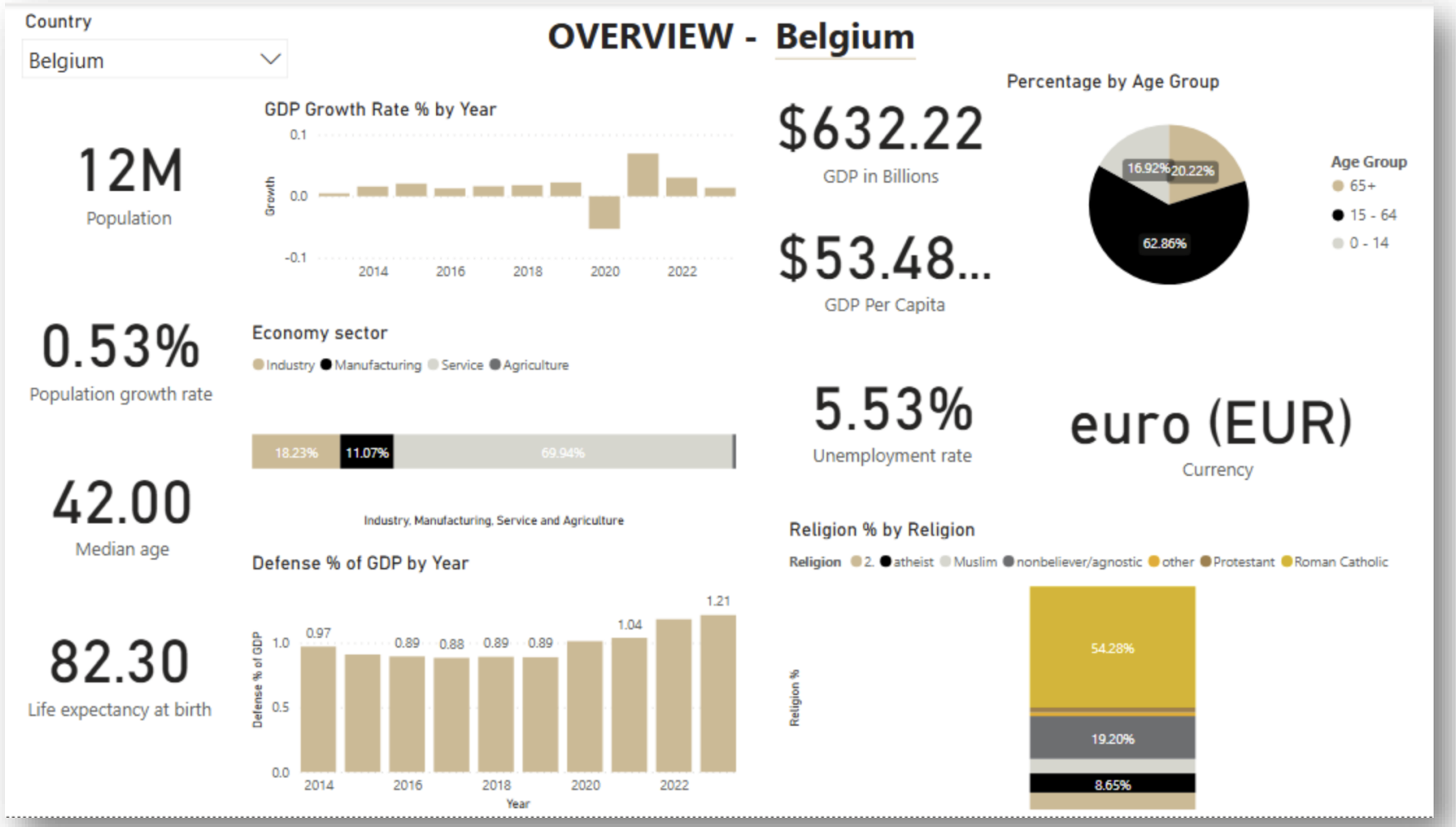
DATA PROCESSING + AUTOMATION

We created a script that centralizes the running of code for ease of use

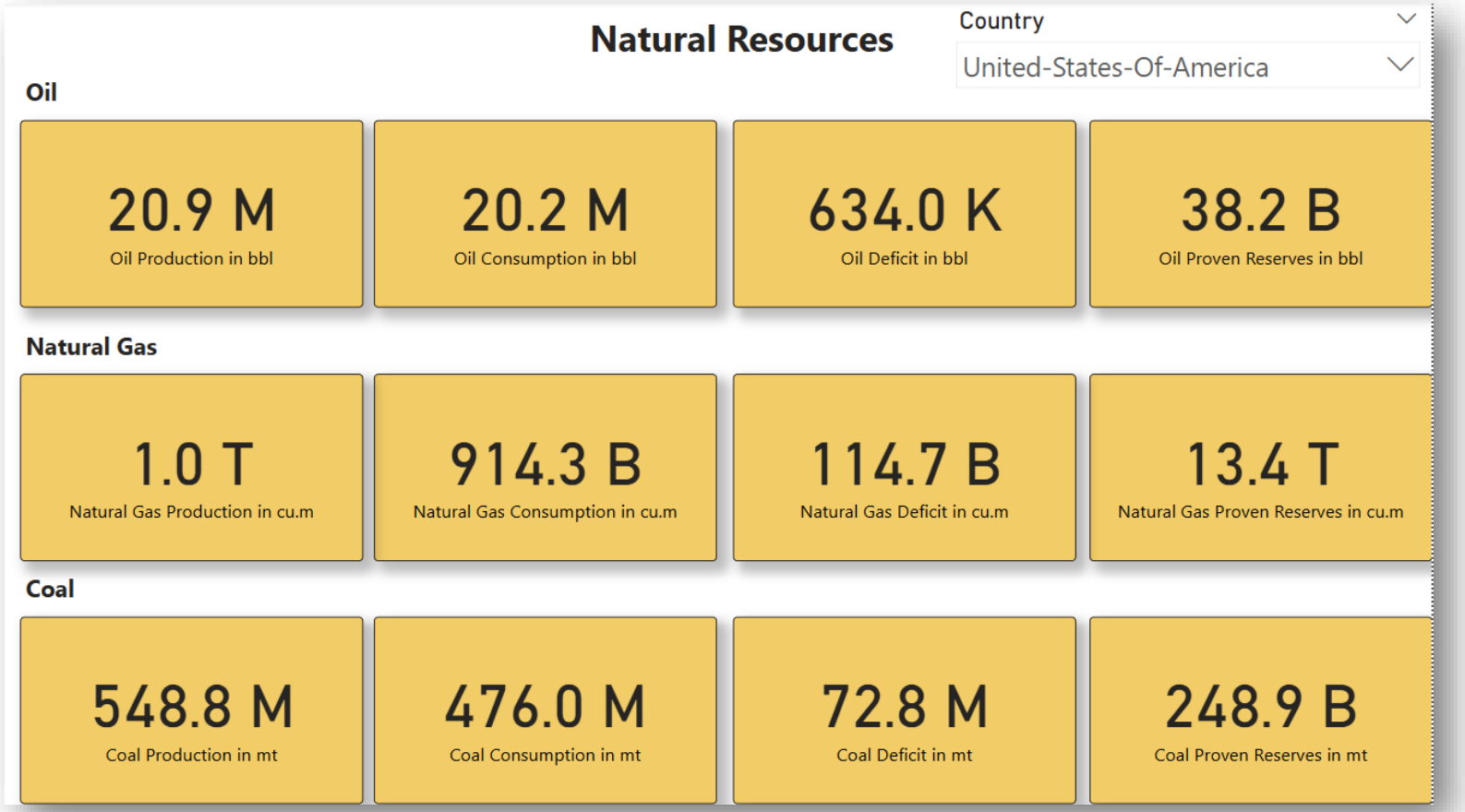
```
#!/bin/bash

scripts=("Part_1/Overview/scrape_overview_cia.py" "Part_1/Overview/world_bank_scrape.py" "Part_1/Natural

for script in "${scripts[@]}"; do
    echo "$(date +'%Y-%m-%d %H:%M:%S') - Running: $script..."
    python3 "$script"
    echo "$(date +'%Y-%m-%d %H:%M:%S') - Finished: $script ✅"
    echo "--------------------------------"
done

echo "$(date +'%Y-%m-%d %H:%M:%S') - All scripts executed successfully! 🎉"
```

DASHBOARDS

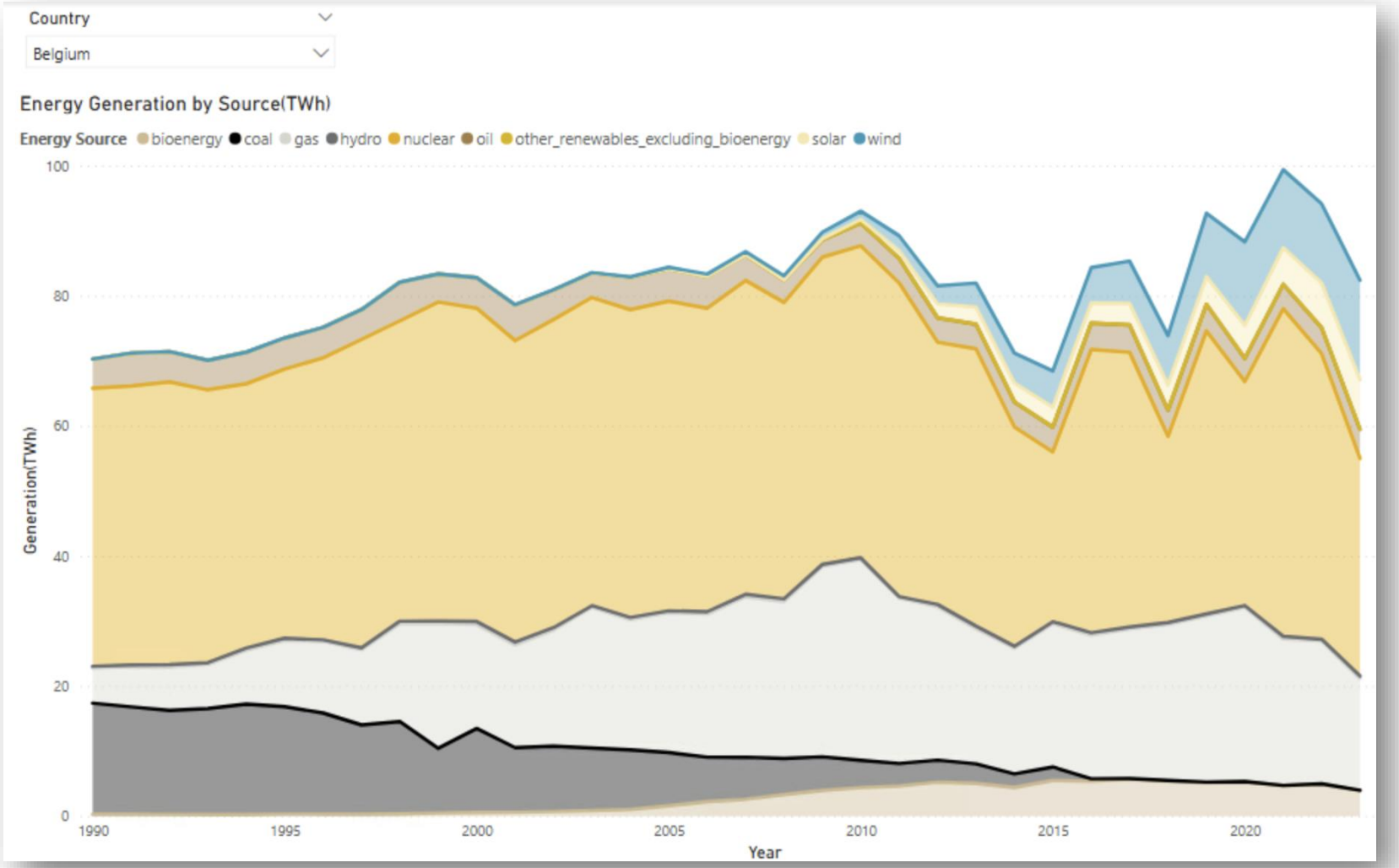
OVERVIEW


Country
Belgium
OVERVIEW - Belgium
12M
Population
0.53%
Population growth rate
42.00
Median age
82.30
Life expectancy at birth
GDP Growth Rate % by Year
0.1
0.0
-0.1
Growth
2014
2016
2018
2020
2022
Economy sector
Industry
Manufacturing
Service
Agriculture
18.23%
11.07%
69.94%
Industry, Manufacturing, Service and Agriculture
Defense % of GDP by Year
0.97
0.89
0.88
0.89
0.89
1.04
1.21
1.0
0.5
0.0
Defense % of GDP
2014
2016
2018
2020
2022
Year
$632.22
GDP in Billions
$53.48...
GDP Per Capita
5.53%
Unemployment rate
euro (EUR)
Currency
Percentage by Age Group
16.92%
20.22%
62.86%
Age Group
65+
15 - 64
0 - 14
Religion % by Religion
Religion
2.
atheist
Muslim
nonbeliever/agnostic
other
Protestant
Roman Catholic
Religion %
54.28%
19.20%
8.65%

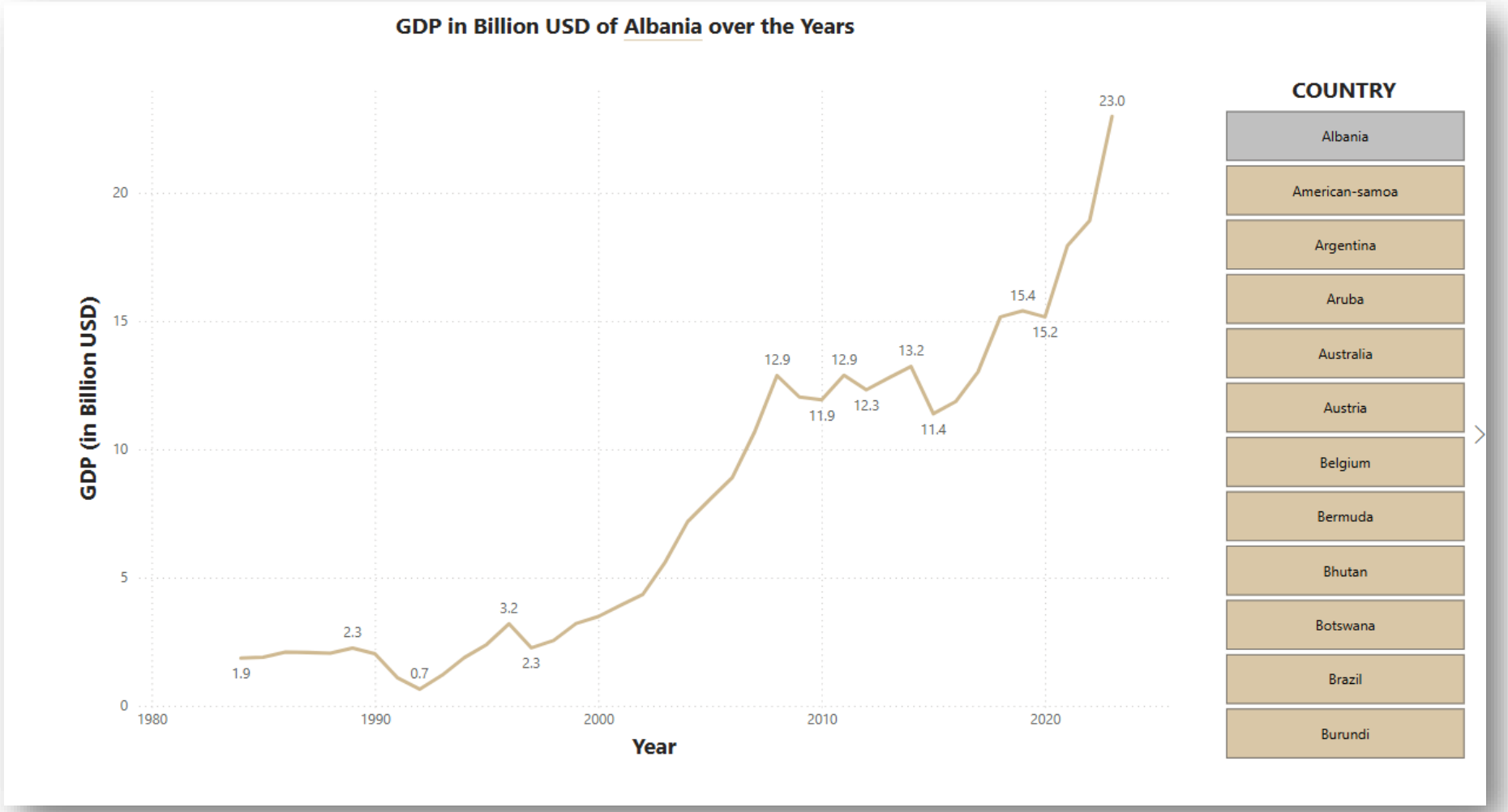
Natural Resources



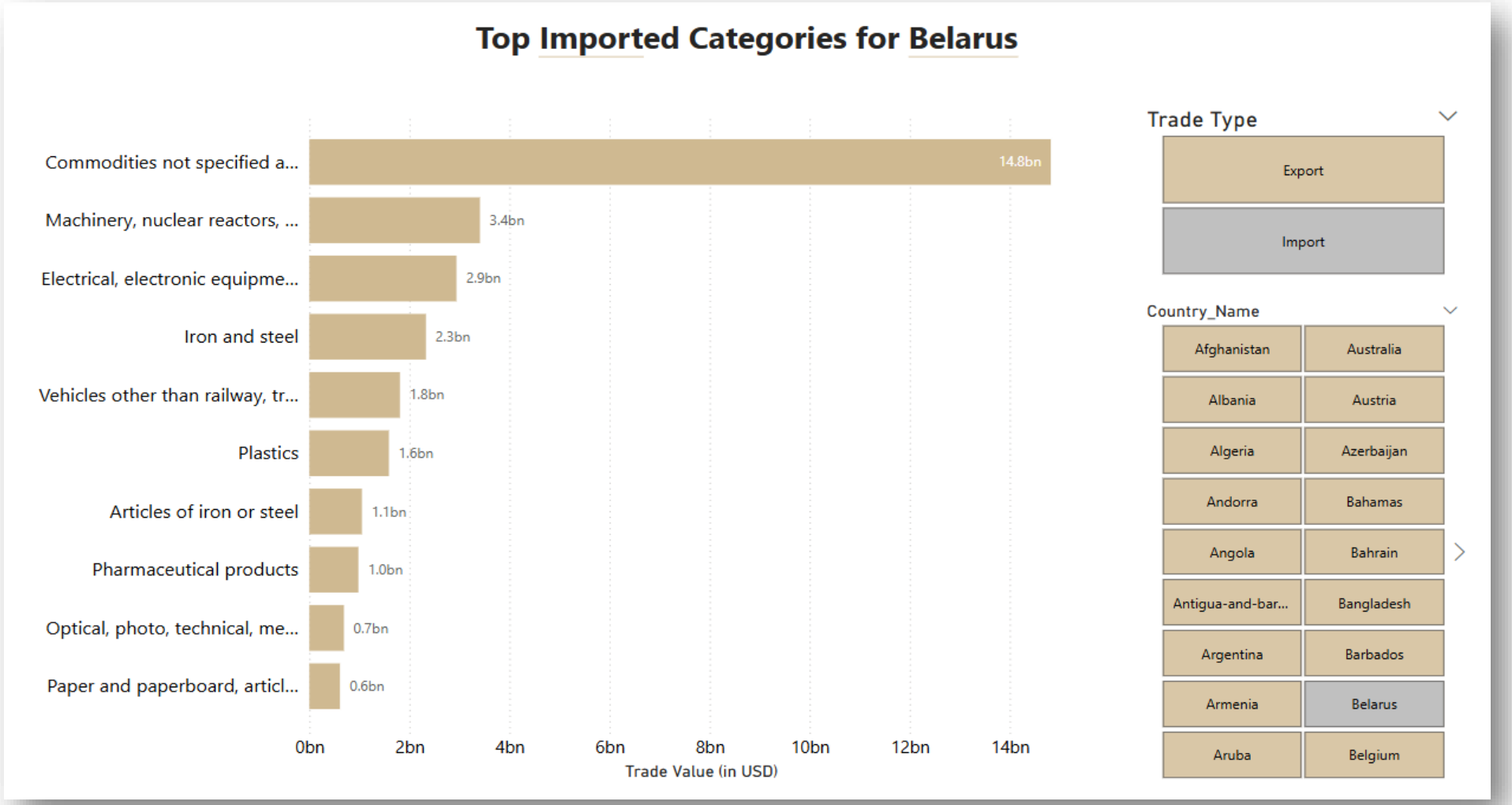
Energy


Country
Belgium
Energy Generation by Source(TWh)
Energy Source bioenergy coal gas hydro nuclear oil other_renewables_excluding_bioenergy solar wind
100
80
60
40
20
0
Generation(TWh)
1990
1995
2000
2005
2010
2015
2020
Year

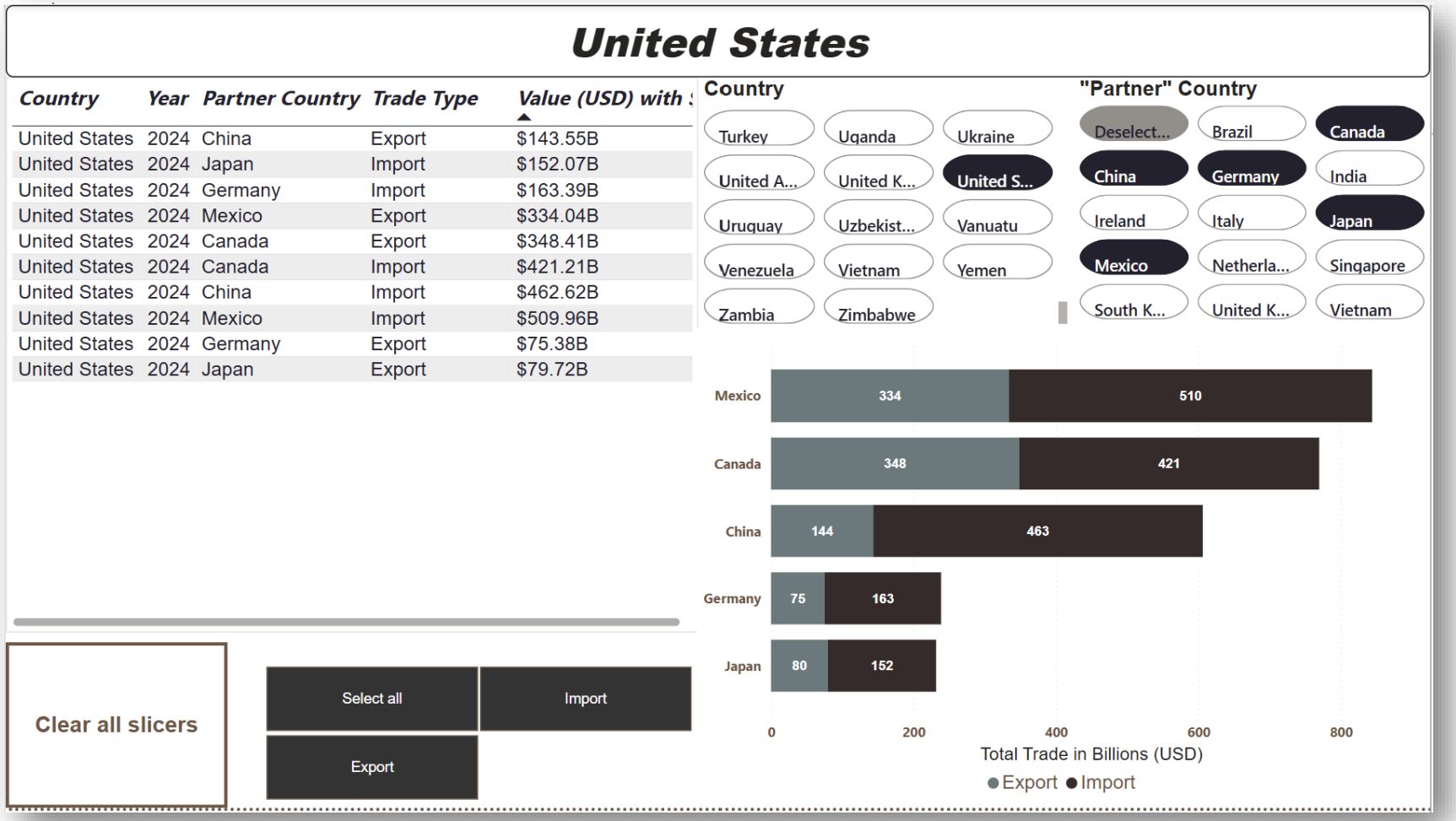
ECONOMICS - GDP



TRADE BY CATEGORY

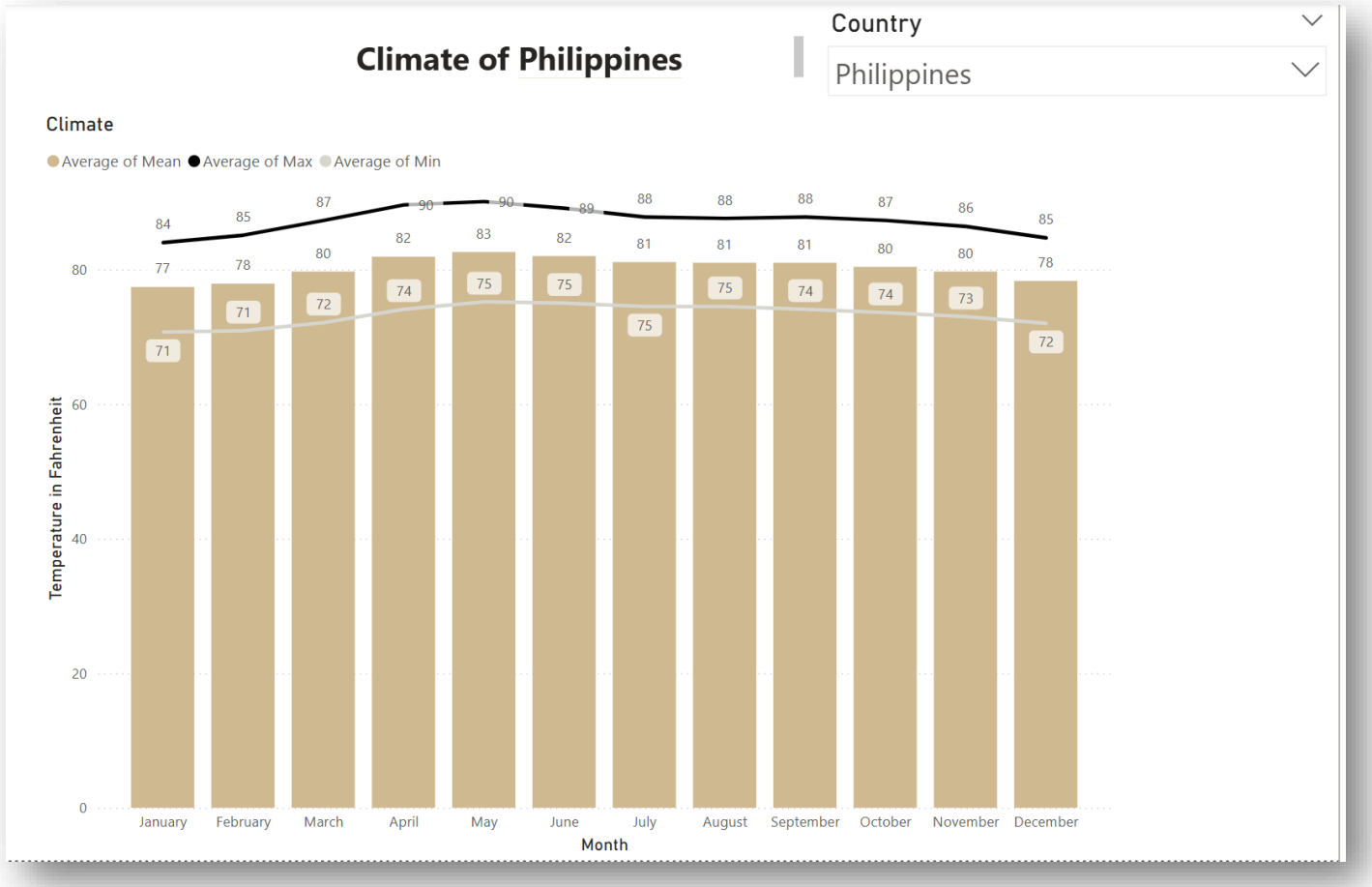


TRADE BY COUNTRIES



PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Climate


Climate of Philippines
Country
Philippines
Climate
Average of Mean
Average of Max
Average of Min
Temperature in Fahrenheit
Month
January
February
March
April
May
June
July
August
September
October
November
December

GEOGRAPHY

United States

Country

United States

Land Area Sq. km	Water Area Sq. km
9147593	685924

Total Area Sq. km	Mean Elevation
10M	760

Highest Point (m)	Lowest Point (m)
Denali (Mount McKinley)	Badwater Basin (Death Valley)

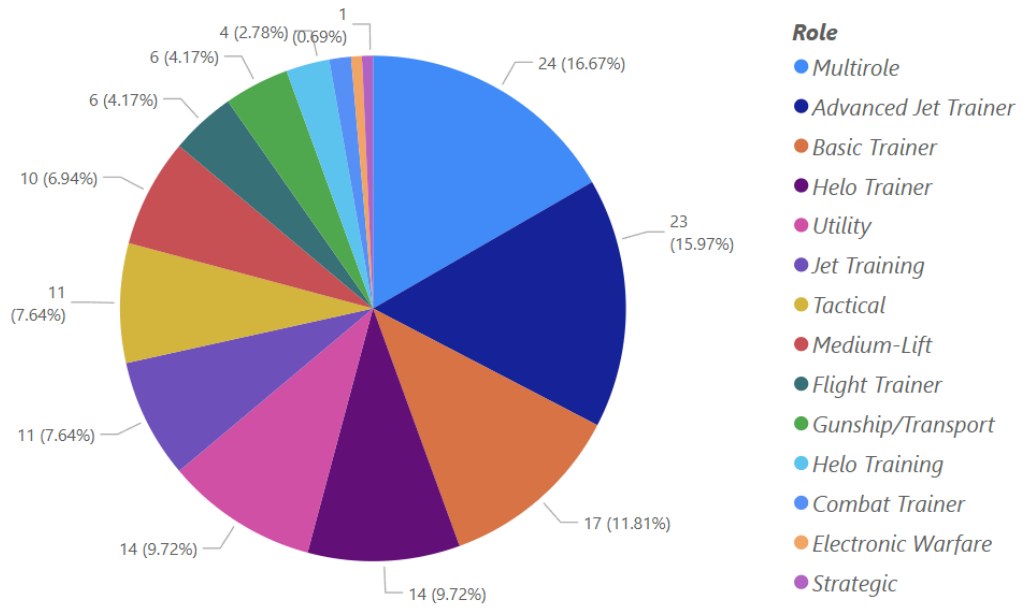
Border Length	Bordering Countries
12002	Canada 8,891 km (including 2,475 km with Alaska); Mexico 3,111 km

MILITARY – AIR POWER

Country

Venezuela

Aircraft	Role	Sum of Units
AS532	Medium-Lift	10
Boeing 707	Strategic	1
C-130H	Tactical	3
Cessna 208	Utility	3
DA42	Flight Trainer	6
Do 228NG	Utility	3
EMB312	Basic Trainer	17
Enstrom 280	Helo Trainer	2
Enstrom 480	Helo Trainer	12
Enstrom 480	Helo Training	4
F-16A	Multirole	3
F-16B	Combat Trainer	2
JL-8	Advanced Jet Trainer	23
King Air 200/350	Utility	5
Metroliner III	Electronic Warfare	1
Metroliner IV	Utility	1
Mi-17	Gunship/Transport	6
SF.260	Jet Training	11
Short 360	Utility	2
Su-30MK2	Multirole	21
Y-8	Tactical	8
Total		**144**

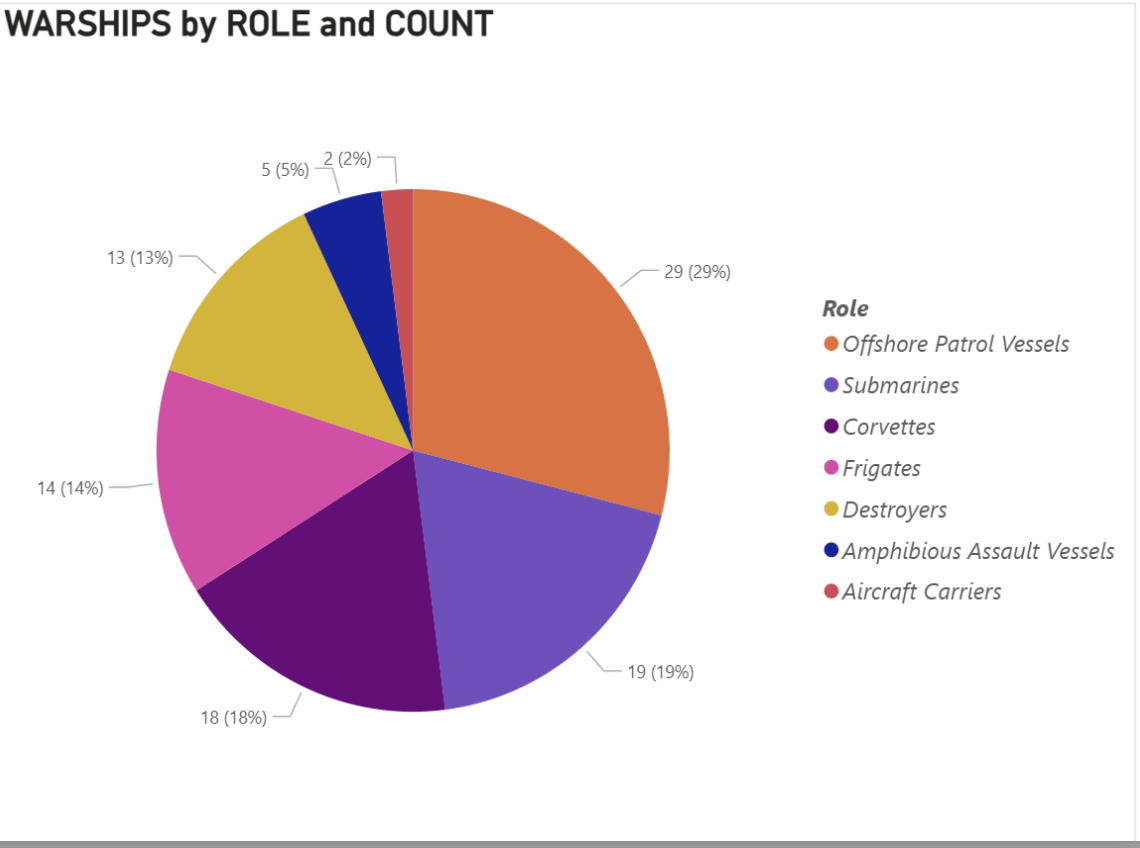


MILITARY – NAVAL POWER

Country

India

Country	Role	Ship	Sum of Unit
India	Aircraft Carriers	Vikramaditya	
India	Aircraft Carriers	Vikrant	
India	Amphibious Assault Vessels	Austin-class	
India	Amphibious Assault Vessels	Magar-class	
India	Amphibious Assault Vessels	Shardul-class	
India	Corvettes	Abhay-class	
India	Corvettes	Kamorta-class	
India	Corvettes	Khukri-class	
India	Corvettes	Kora-class	
India	Corvettes	Veer-class	
India	Destroyers	Delhi-class	
India	Destroyers	Kolkata-class	
India	Destroyers	Rajput-class	
India	Destroyers	Visakhapatnam-class	
India	Frigates	Brahmaputra	
India	Frigates	Nilgiri-class	
India	Frigates	Shivalik-class	
India	Frigates	Talwar-class	
India	Offshore Patrol Vessels	Bangaram-class	
India	Offshore Patrol Vessels	Car Nicobar-class	1
India	Offshore Patrol Vessels	Saryu-class	
India	Offshore Patrol Vessels	Sukanya-class	
Total			**10**



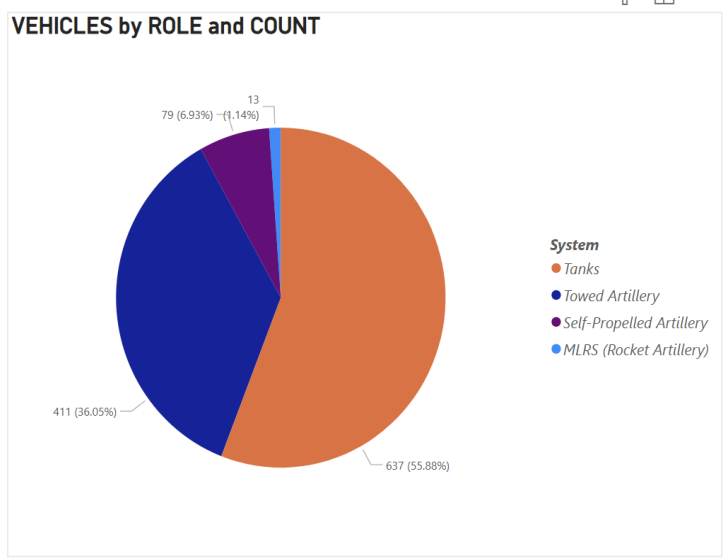
MILITARY – LAND POWER

Country

Chile

Chile

Country	Metric	System	Sum of Value
Chile	Readiness	MLRS (Rocket Artillery)	5
Chile	Stock	MLRS (Rocket Artillery)	8
Chile	Readiness	Self-Propelled Artillery	31
Chile	Stock	Self-Propelled Artillery	48
Chile	Readiness	Tanks	251
Chile	Stock	Tanks	386
Chile	Readiness	Towed Artillery	162
Chile	Stock	Towed Artillery	249
Chile	Readiness	Vehicles	5741
Chile	Stock	Vehicles	8832
Total			**15713**



Country

Chile

Role	Sum of System Name	Sum of Year
105mm Pack Howitzer / Lightweight Artillery Piece.	25	1956
155mm Self-Propelled Artillery (SPA)	20	1963
155mm Towed Field Gun	42	1917
155mm Towed Howitzer	8	1975
37mm Anti-Tank Gun	37	1940
4x4 High-Mobility Multi-Purpose Wheeled Vehicle	2	1985
4x4 Multi-Purpose Light Utility Vehicle	23	1960
4x4 Utility Truck	30	1947
6x6 Wheeled Armored Car	10	1974
6x6 Wheeled Multirole Armored Fighting Vehicle (AFV)	9	1974
8x8 Wheeled Armored Fighting Vehicle (AFV)	13	1972
8x8 Wheeled Armored Reconnaissance Vehicle	5	1983
Amphibious Armored Personnel Carrier	24	1956
Four-Wheeled Armored Car	38	1940
Half-Track Multi-Purpose Armored Personnel Carrier	36	1940
Infantry Fighting Vehicle (IFV)	7	1975
Light Armored Reconnaissance Vehicle	12	1973
Light Tank (LT)	61	3893
Light Tank (LT) Tracked Combat Vehicle	28	1951
Main Battle Tank (MBT)	43	5910
Medium Tank	32	1942
Medium Tank Tracked Combat Vehicle	29	1951
Total	**903**	**82281**

Image URL

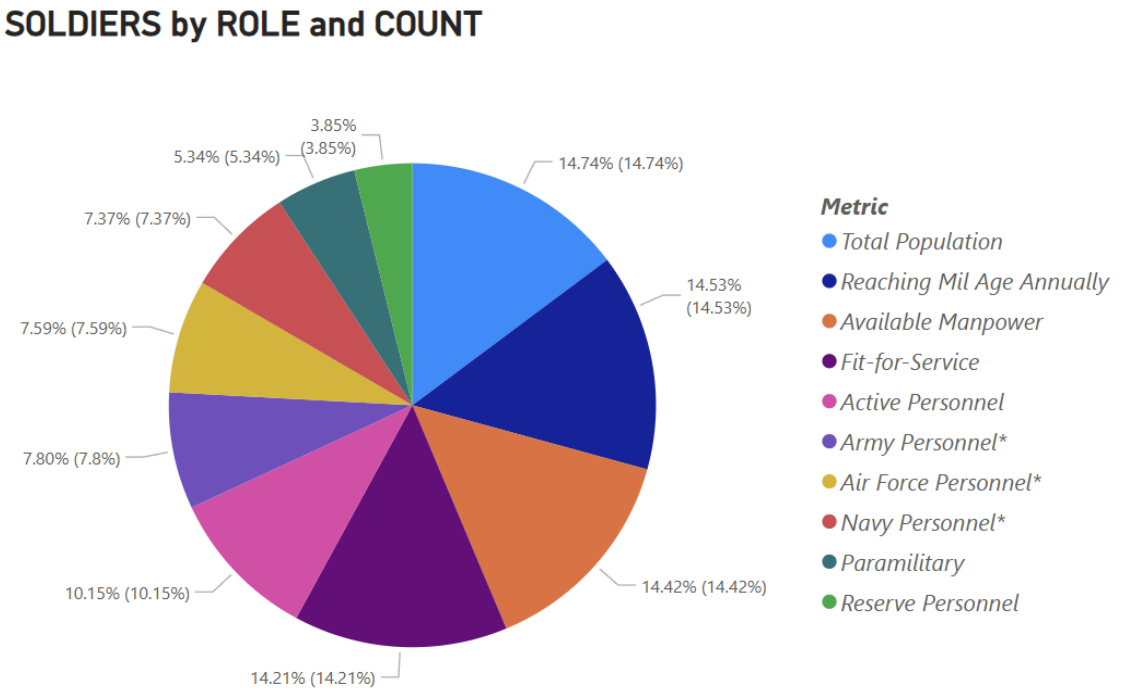
MILITARY – MAN POWER

Country

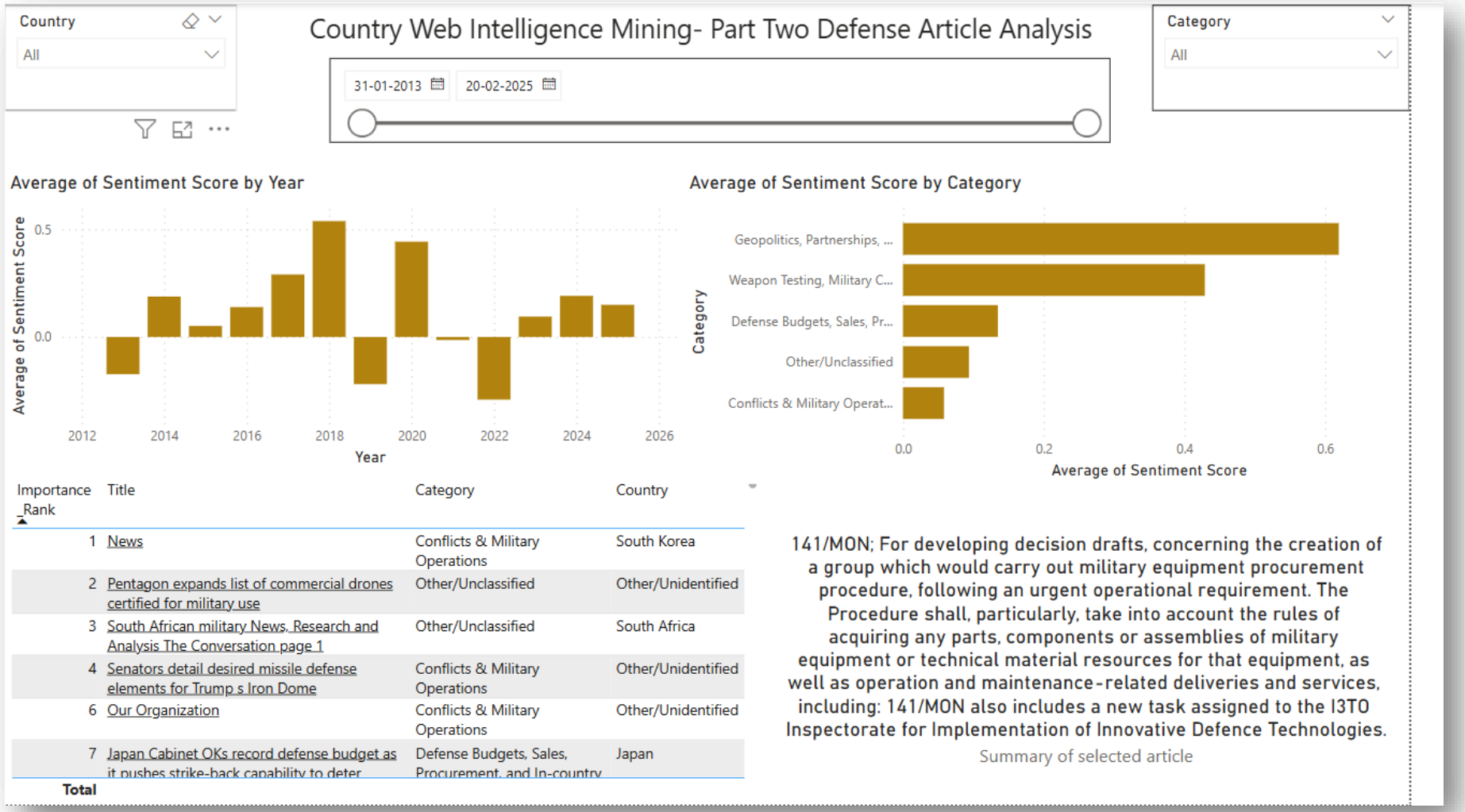
Bahrain

Bahrain

Country	Metric	Sum of Value
Bahrain	Active Personnel	95
Bahrain	Air Force Personnel*	71
Bahrain	Army Personnel*	73
Bahrain	Available Manpower	135
Bahrain	Fit-for-Service	133
Bahrain	Navy Personnel*	69
Bahrain	Paramilitary	50
Bahrain	Reaching Mil Age Annually	136
Bahrain	Reserve Personnel	36
Bahrain	Total Population	138
Bahrain	Yearly Mobilization Potential	17018
Total		**17954**



Part 2- DEFENSE ARTICLE ANALYSIS



Overview of Libraries & Functions Used

Libraries Used

- **Web Scraping & Data Retrieval:** requests, BeautifulSoup, newspaper3k, serpapi
- **Natural Language Processing:** nltk, transformers, VADER, sumy, re
- **Data Processing & ML:** pandas, numpy, scikit-learn (TfidfVectorizer)
- Will require an API key for google scraping which is free for the most part but if multiple pulls per day are needed it costs around 150USD per month for 15000

Key Functions & Their Role

- **Data Collection:** fetch_page(), scrape_articles(), search_google_articles() – Ensures automated & diverse data collection.
- **Text Processing:** clean_text(), extract_article_content(), summarize_article() – Prepares clean, structured text.
- **Sentiment & Stance Analysis:** analyze_sentiment(), analyze_stance() – Determines tone & positioning of articles.
- **Categorization & Country Mapping:** classify_category(), identify_country() – Segments articles by relevance.

Ranking and Importance Scoring

Ranking Method: TF-IDF

Why TF-IDF?

- Relevance-Based: Identifies key terms & informative content.
- Context-Aware: Filters out common words to highlight unique insights.
- Scalable: Efficient for large datasets without manual intervention.

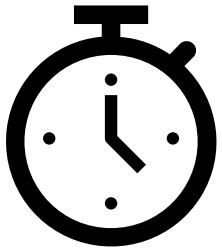
How It Works

- TF (Term Frequency): Measures word occurrence in an article.
- IDF (Inverse Document Frequency): Reduces weight of common words.
- Final Score: Sum of TF-IDF values in the article's summary.

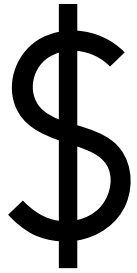
Why Not Sentiment Scores?

- Sentiment ≠ Importance (e.g., neutral news on military conflict could be more critical than a positive minor update).
- TF-IDF focuses on information content, not emotional tone.

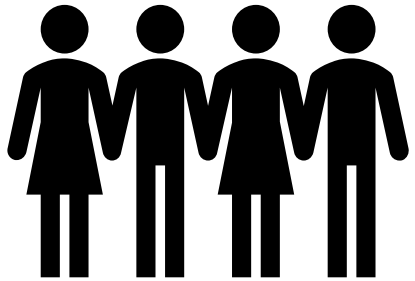
ACTIONABLE RESULTS



- Saves **10+ hours** per week of labor-intensive scraping.



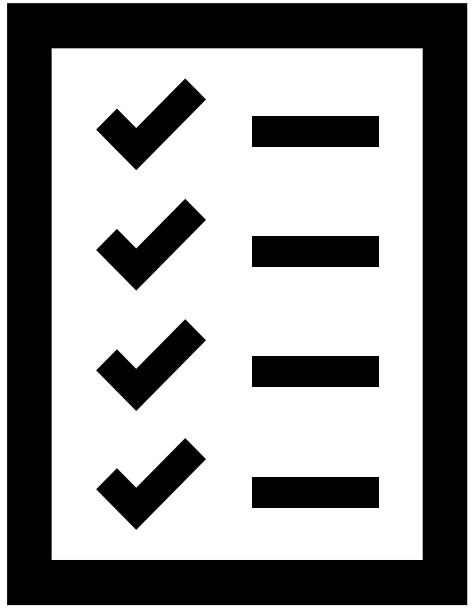
- On average, these improvements will save **$2000 +** per month.



- This will free up **4-5 data analysts** to work on other tasks.

SUMMARY

1. Business Problem
2. Assumptions
3. Approach + Decisions Taken
4. Code Operations + Outputs
5. Processing and Automation
6. PowerBI Dashboard Outputs Demo



THANK YOU

Questions?

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business