# Understanding Wildlife Trafficking using Big Data Analysis Techniques

ASHUTOSH KUMAR* and JAYANTH ANALA†*, New York University, USA

The D-ISN Project through Advanced Data Analysis and ETL Pipeline is the base for this project. In this project, we aimed at combating the global issue of wildlife trafficking, by making significant improvements through advanced data analysis techniques, visualization strategies, and the augmentation of the existing ETL pipeline and it's scalability. Recognizing the multifaceted impact of illegal wildlife trade on society, the project seeks to leverage data-driven insights to refine disruption techniques and contribute to the broader understanding of illicit supply networks.

Additional Key Words and Phrases: ETL, Big Data, Wildlife, PySpark, HDFS

**ACM Reference Format:**

Ashutosh Kumar and Jayanth Anala. 2023. Understanding Wildlife Trafficking using Big Data Analysis Techniques. In . ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The illegal wildlife trade and trafficking represent critical global challenges that extend beyond economic, safety, and conservation concerns. These illicit activities not only contribute to animal abuse, injuries, and fatalities but also pose significant public health risks, as exemplified by the potential spread of zoonotic diseases like the COVID-19 virus through animals. Effectively addressing and mitigating the impact of illegal wildlife trade requires a comprehensive understanding of the underlying patterns and dynamics. Leveraging the capabilities of the Big Data ecosystem becomes imperative in this context, given its capacity to store, manipulate, analyze, and mine vast datasets related to these trades.

By harnessing the power of Big Data analytics, decision-makers can gain valuable insights into the intricate networks and mechanisms driving illegal wildlife trade. Through thorough data cleaning and exploration methods, we can uncover hidden patterns, identify key players, and understand the modus operandi of traffickers. This knowledge creation is pivotal in formulating informed strategies to prevent, monitor, and combat illegal wildlife trade effectively.

Our approach involves correlating datasets encompassing wildlife advertisements with publicly available wildlife information. We correlated the wildlife advertisements data and wildlife-information datasets available to public [1]. This interdisciplinary analysis provides a holistic perspective, enabling us to connect the dots between trade activities and their impact on wildlife populations. By merging these datasets, we aim to contribute to a more comprehensive understanding of the factors influencing illegal wildlife trade, ultimately empowering stakeholders to make informed decisions and implement targeted interventions in the ongoing fight against this global issue.

*Both authors contributed equally to this research.
†Github link: https://github.com/jayanthanala/BigData


Manuscript submitted to ACM

## 2 PROBLEM FORMULATION

### 2.1 Problem Identification

During the analysis of the dataset on Wildlife Trafficking Advertisements obtained through the ETL pipeline, several problems were identified, particularly related to the identification and extraction of relevant data. The ETL pipeline seems to be crawling advertisements that are unrelated to the specified case study. Examples range from products like "Nike Air Jordan 1 Animal Print" to "1990 Cat Playing Card" indicating a issue in the criteria used to identify related ads.

Upon closer inspection, it is easy to see that many columns, apart from the description of the advert or product, contain null or empty values. This lack of information poses a significant obstacle in deriving meaningful insights from the dataset. The absence of data in key fields effects the ability to process, clean, and validate the dataset accurately, making it challenging to extract reliable statistical answers, identify patterns, and generate insightful visualizations.

To address this challenge, a comprehensive review of the ETL pipeline's criteria for identifying relevant advertisements is necessary. It is crucial to refine the parameters or enhance the algorithms used in the extraction process to ensure that only data directly related to wildlife trafficking is captured. Additionally, efforts should be directed toward improving data accuracy and completeness by encouraging advertisers to provide more comprehensive information.

As part of our methodology, we are exploring the correlation of this dataset with additional datasets from other sources. This integration is aimed at providing a more comprehensive understanding of the wildlife trafficking landscape. However, the success of this correlation relies heavily on addressing the issues of inaccurate identification and incomplete data within the primary dataset. By enhancing the quality and relevance of the crawled data, we can ensure that our statistical analyses, pattern identifications, and visualizations can show accurate insights for the specified case scenario.

### 2.2 Challanges & Constraints

- To perform the data exploration, cleaning the advertisement datasets which contains lots of unwanted and unnecessary data is required.
- The title, product, text and description are all basically same columns some being null. It contains all the metadata of the website the advt is posted which is lot of data/strings to process.
- Data on family, genus, taxa-names of the species were out-of-scope for the advt dataset. Comman names of the animal were supposed to extracted from the extra long strings which includes metadata of site.
- Finding datasets of wildlife trafficking was difficult due to lack of information. We could find few datasets mentioned in the next sections, but those require approval for data extraction and download. We got approval for few but few has approval on unit levels.

## 3 RELATED WORK

### 3.1 PySpark

PySpark is the Python library for Apache Spark, an open-source, distributed computing system designed for big data processing and analytics. PySpark provides a high-level interface that allows developers to write data-intensive applications in Python, leveraging the power of Spark's distributed computing capabilities. At its core, PySpark facilitates the seamless integration of Python with Spark's resilient distributed datasets (RDDs), dataframes, and machine learning libraries. With its ability to scale horizontally across a cluster of machines, PySpark enables the processing of large-scale datasets in a parallel and fault-tolerant manner. This combination of Python's simplicity and Spark's distributed

computing prowess makes PySpark a versatile and efficient tool for handling diverse data processing tasks, including data cleaning, transformation, analysis, and machine learning on massive datasets.

### 3.2 Tableau

Tableau is commonly used for Exploratory Data Analysis (EDA) due to its powerful and intuitive data visualization capabilities. It allows users to interactively explore and analyze complex datasets, uncover patterns, and gain insights through visually appealing charts, graphs, and dashboards. Tableau's drag-and-drop interface and wide range of visualization options make it accessible to users with varying levels of technical expertise. By providing a dynamic and interactive environment, Tableau facilitates a deeper understanding of data relationships, trends, and outliers. Its real-time interactivity allows analysts to quickly iterate and explore different angles of the data, fostering a more efficient and effective EDA process. Overall, Tableau's visualization features make it a popular choice for transforming raw data into meaningful insights, aiding in decision-making and communication of findings.

### 3.3 Literature Review

We've examined the foundational paper that outlines the process for obtaining the dataset. This pipeline was designed with the primary goal of filling in data gaps by allowing the generation of comprehensive and varied datasets. These datasets encompass a broad spectrum of animal species and diverse web-market locations. Our dataset, as far as we know, is groundbreaking as it is the first to encompass a wide array of endangered species being promoted across multiple countries[2]. This expansion and inclusion of various endangered species and global locations enhance the dataset's richness and contribute to a more comprehensive understanding of the advertising landscape for these animals.

Some studies and reports describe traffickers' modus operandi, or factors that may influence their decisions to traffic products through certain ports over others[1].A key area of concern in combating WT is the convergence of multiple forms of illicit trade. Convergence can take a variety of forms. For instance, revenue from WT activities can fund arms trafficking. Seizure data provides a glimpse of how WT networks operate, alert experts to trends in supply and demand for different species, and point to key locations for deterring wildlife crime

## 4 METHODS, ARCHITECTURE AND DESIGN

Our main objective was to explore the Advert dataset by connecting it with a detailed list of wildlife types associated with illegal wildlife trade (IWT) (i.e.,wildlife seizures) seizures [2]. We also wanted to link this information with the purposes for which these wildlife items were being traded (i.e., use-type). The additional dataset we used for this purpose includes data from 2010, as trading wildlife online started gaining traction around that time, especially on the internet and dark web. By combining these datasets, we aimed to gain insights into the types of wildlife being traded illegally and understand the reasons behind these activities.

The Disrupting Operations of Illicit Supply Networks (D-ISN) project, focused on combating the illegal trade in wild animals, should be understood initially for a major upgrade. The plan involves using advanced data analysis techniques, impactful visualizations, and proposing a new improvements the existing Extract, Transform, Load (ETL) pipeline. Knwoing the urgent need for innovative solutions to tackle wildlife trafficking challenges, we adopted cutting-edge data analysis methods. Techniques like exploratory data analysis and quantitative analysis will help uncover patterns, trends, and hidden connections in the data. The project is also placing a strong emphasis on visualization to make complex information easily understandable. Visual tools, like trade locations and social network analysis visualizations, will be used to showcase research findings. These visuals will not only help in understanding the scale and nature of illicit

wildlife trade but will also assist in effectively communicating results to diverse audiences. The analysed data from D-ISN project aims to make significant progress in combating wildlife trafficking by understand the patters and insights using advanced data analysis techniques, and impact visualizations. These enhancements can deepen our understanding of illicit supply networks and empower stakeholders with actionable insights for informed decision-making in addressing this global challenge.

### 4.1 Datasets Gathering

We have few datasets relating to the trade details and species details where the columns merge using various features like taxa-name, common name of the species. We were given datasets split into multiple parts in parquet format with each part containing around 2000 records. We have used Pyspark to merge all the parquet files into a single data-format to perform cleaning and analysis techniques. It came out to be 375726 record in total. The datasets from other sources are listed as follows

- CITES trade database, using their website (https://trade.cites.org/)
- The TRAFFIC International Wildlife Trade Portal, using their website (https://www.wildlifetradeportal.org/)
- (IWT)Dataset of seized wildlife and their intended uses (https://doi.org/10.6084/m9.figshare.14914773.v1)
- World Wildlife Crime Report (https://www.unodc.org/unodc/en/data-and-analysis/wildlife.html)

We will be using 2 datasets from the above list, the remaining could not be used due to restrictions or required-approval on unit levels to extract and download the data.

### 4.2 External Dataset Selection

We identified four external datasets to enrich our analysis of the advertisement dataset, seeking to extract meaningful insights. One of the datasets, from TRAFFIC, required approval for access, which we successfully obtained. However, it introduced a level of complexity, as approval was needed for every query or use of the dataset. Despite this hurdle, the TRAFFIC dataset promises valuable information, particularly in terms of common names associated with wildlife, which aligns with our goal of identifying animal names in advertisements.

The World Wildlife Crime Report dataset from UNODC was relatively small, containing data only for the year 2023. While its size might limit the scope of insights we can derive, it still holds relevance for understanding recent trends and occurrences in wildlife crime, providing a temporal context to our analysis.

The CITES trade database, focusing on legal transactions of wildlife, offers a contrasting perspective to our primarily illicit wildlife trade dataset. As it is freely accessible to the public, we derived a snapshot of this dataset for analysis. Notable information includes taxa names, family, genus, common names, and the location of trade, providing valuable geo-location context to complement our study.

Lastly, the Seized Wildlife and their intended uses dataset appears to be a rich source of reusable data. With features like taxa names, family, genus, use type, class, kingdom, species, and order, this dataset aligns well with our objectives. Merging this dataset with our advertisement data may reveal insightful patterns, shedding light on the intended uses of seized wildlife and facilitating a deeper understanding of the motivations behind illegal wildlife trade.

By exploring and integrating these external datasets, we aim to gain a more comprehensive understanding of the wildlife trade landscape, encompassing both legal and illegal aspects, and derive actionable insights to address the challenges posed by such activities.

Table 1. Use-type subcategories with number of taxa in each subcategory

| Use-type(main) | Use-type(subcategory) | Number-of-taxa |
|---|---|---|
| 7262 | 9714 | 99161 |
| live | live | 2,173 |
| dead/raw | dead (whole animal) | 1,642 |
| dead/raw | animal parts (bone or bone-like) | 623 |
| dead/raw | animal fibers | 445 |
| unspecified | unspecified | 426 |
| dead/raw | skin/leather (raw) | 414 |
| dead/raw | food (raw) | 389 |
| dead/raw | taxidermy | 381 |
| dead/raw | animal parts (fleshy) | 297 |
| processed/derived | clothing | 262 |
| processed/derived | skin/leather (products) | 258 |
| dead/raw | shells (raw) | 233 |
| processed/derived | medicine | 216 |
| processed/derived | derivative | 209 |
| processed/derived | jewellery & personal ornaments | 203 |
| dead/raw | coral (dead) | 174 |
| dead/raw | wood/timber | 163 |
| dead/raw | extract | 147 |
| processed/derived | carvings/engravings | 126 |
| processed/derived | shells (product) | 123 |

Table 2. Total Null/Invalid Values in Dataset w.r.t Columns

| title | text | name | description | category | price | currency | seller | location | country | lat | lon |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7262 | 9714 | 99161 | 101145 | 367803 | 181097 | 183221 | 355721 | 351265 | 375233 | 375233 | 375233 |

### 4.3 Data Pre-processing and Cleaning

In our data pre-processing, the first step is to eliminate the records which are invalid. To find the valid records we have certain columns required such as title, name, description, price, label_product and product. In return, if we are unable to find the important attributes like location and species name in the dataset we consider them as useless records.

Looking at Table 1, we noticed a bunch of missing values in each record, indicating inconsistency in the data. Specifically, we identified 9,714 records lacking a description and 375,233 records with no country or location details, presenting a significant challenge. Instead of discarding these values, we opted to find the species name by leveraging the available information from descriptions and product names. During our analysis, we noted that the "ships to" column had invalid entries in many instances, necessitating a syntax fix in the pipeline. We also observed some unnecessary characters in the title column, and we took steps to correct those. The join operation posed a challenge since it was a one-to-many join due to multiple records in trade datasets when the trade involved a legal purchase.

We noticed that the trade databases employed slightly different terminology for use-types. To address this inconsistency, we standardized and unified the use-types across the three trade databases. This involved ensuring that similar concepts were represented with consistent wording. Additionally, we took the initiative to create 'Internet-friendly'

```
|-- url: string (nullable = true)
|-- title: string (nullable = true)
|-- text: string (nullable = true)
|-- domain: string (nullable = true)
|-- name: string (nullable = true)
|-- description: string (nullable = true)
|-- image: string (nullable = true)
|-- retrieved: string (nullable = true)
|-- production_data: string (nullable = true)
|-- category: string (nullable = true)
|-- price: double (nullable = true)
|-- currency: string (nullable = true)
|-- seller: string (nullable = true)
|-- seller_type: string (nullable = true)
|-- seller_url: string (nullable = true)
|-- location: string (nullable = true)
|-- ships_to: string (nullable = true)
|-- id: string (nullable = true)
|-- loc_name: string (nullable = true)
|-- lat: double (nullable = true)
|-- lon: double (nullable = true)
|-- country: string (nullable = true)
|-- product: string (nullable = true)
|-- label_product: string (nullable = true)
|-- score_product: double (nullable = true)
|-- label: double (nullable = true)
|-- score: double (nullable = true)
```

Fig. 1. Columns in Wildlife Trafficking - Advertisement Dataset

search terms associated with each use-type. These search words served as alternatives or synonyms, making it easier to query and retrieve relevant information. For instance, in the case of the use-type "foetus," we generated search words such as "foetus," "fetus," "placenta," and "embryo" to cover various terms that might be used in different databases. However, for the "live" and "dead" use-types, we did not assign specific search words. It's worth noting that we refrained from recording the number of occurrences for each taxa-use combination. This decision was influenced by the likelihood of duplicated seizure records across the three trade databases. By standardizing the use-types and providing user-friendly search terms, we aimed to facilitate a more seamless and consistent exploration of the datasets, ensuring that our analysis captures the breadth of information across the different databases.

To streamline our dataset for this study, we decided to drop several columns that were deemed unnecessary. Columns such as "text," "name," "description," "image," "retrieved," "production_data," "category," "seller_type," "seller_url," "ships_to," "id," "loc_name," "lat," "lon," "product," "country," "score_product," "label," "score" were removed because they either contained null values, weren't useful for our study, or couldn't provide any meaningful information. This cleanup process aimed to enhance the dataset's quality and relevance to our research.

### 4.4 Extracting Common Names

Within the dataset, the "title" column contains the titles of the advertisements, and it's uncertain whether or not they include the common names of animals that are either utilized as resources, trafficked, or incorporated into products. To address this uncertainty, we developed a User Defined Function (UDF) named 'extract_animals'. This UDF takes data from the "title" column as input and employs Python's NLTK word_tokenizer function to tokenize the text into an array of individual words. Subsequently, this array is passed through a regex method that compares each word with the common names in TRAFFIC's dataset.

We collected the common names (i.e., vernacular names) from GBIF, for each taxa resolved to GBIF along with the common names for each upstream taxonomic unit. For example, for Psittacus erithacus, we retrieved the species common name (African Gray Parrot). In some instances, GBIF provided multiple common names per taxonomic unit. The comparison aims to identify if any words in the array correspond to known animal names. If a match is found,

indicating the mention of an animal, the identified name is added to a new column created within the dataset. This entire process is seamlessly integrated into the Spark DataFrame using Spark's 'withColumn()' method, where our UDF is passed as a parameter.

```
animals = spark.read.format('csv').options(header='true',inferschema='true').load('dataset.csv')
animalsArray = parFile3.select("gbif_common_name").toPandas().values.reshape(-1)
animalsArrayUnique = np.unique(animalsArray)

def extract_animals(title):
    matches = re.findall(fr'\b(?:{"|".join(animalsArrayUnique)})\b', title, flags=re.IGNORECASE)
    return matches[0] if matches else None

extract_animal_udf = udf(extract_animals, StringType())
dfFinal = df.withColumn("animal_names", extract_animal_udf(col("title")))
dfFinal.write.parquet("advt_df.parquet")
```

Listing 1. extract_animal UDF in PySpark

By implementing this approach, we enhance our ability to associate advertisements with specific animal species, contributing to a more refined understanding of the dataset and facilitating subsequent analyses related to illegal wildlife trade.

### 4.5 Merging Datasets

After completing the processing, cleaning, and extraction of common animal names from the "title" field in the advertisements and incorporating this information as a new column in our dataset, the subsequent stage involves analyzing the dataset using Tableau. To enhance our dataset, we combine or join the updated dataset with the Illegal Wildlife Trade (IWT) seizures dataset, using the 'db_common_name' as the basis for this union. To ensure data quality and assess correlation, we've generated numerous sketches comparing the two datasets. Once satisfied with the quality, a final dataset is saved as a CSV file with headers included. Subsequently, the analysis phase commences in Tableau, leveraging the enriched dataset to draw insights and patterns that can contribute to a deeper understanding of the relationship between wildlife advertisements and illegal wildlife trade seizures.

### 4.6 Software/Technologies Used

We handled all the tasks related to cleaning, processing, and analyzing data using Python PySpark on NYU's dataproc platform. This platform provided us with extensive computing and storage resources, making collaboration easy. We used python packages like pandas, nltk, re and numpy for our implementations. We adopted version control to collaborate on building data models and analyzing our initial datasets. Instead of using Python libraries like Matplotlib and Seaborn for visualization, we opted for Tableau. This decision allowed us to assess how well the application could handle large datasets. When working with the final, larger dataset, we utilized Tableau to create interactive dashboards and visualizations. This approach significantly aided our understanding of the results and insights we uncovered during the analysis.

## 5 RESULTS

After visualising various kinds of data using various layers and filters we were able to derive interesting patterns and factual information one can only obtain if two datasets are processed and seen together. We have designed interactive

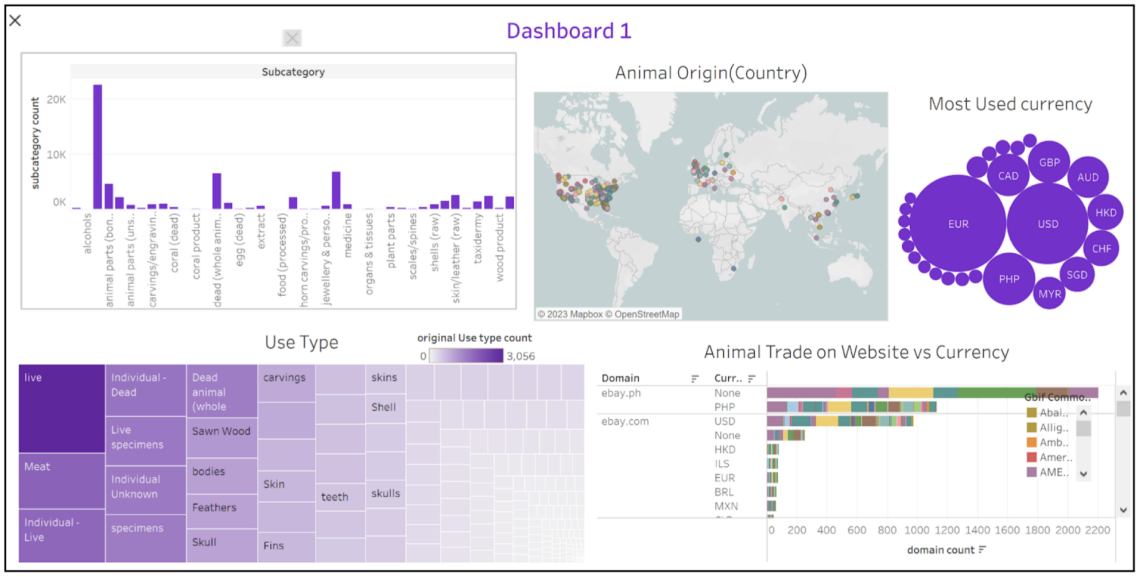


Fig. 2. Interactive Dashboard used to get insights based on various filters.

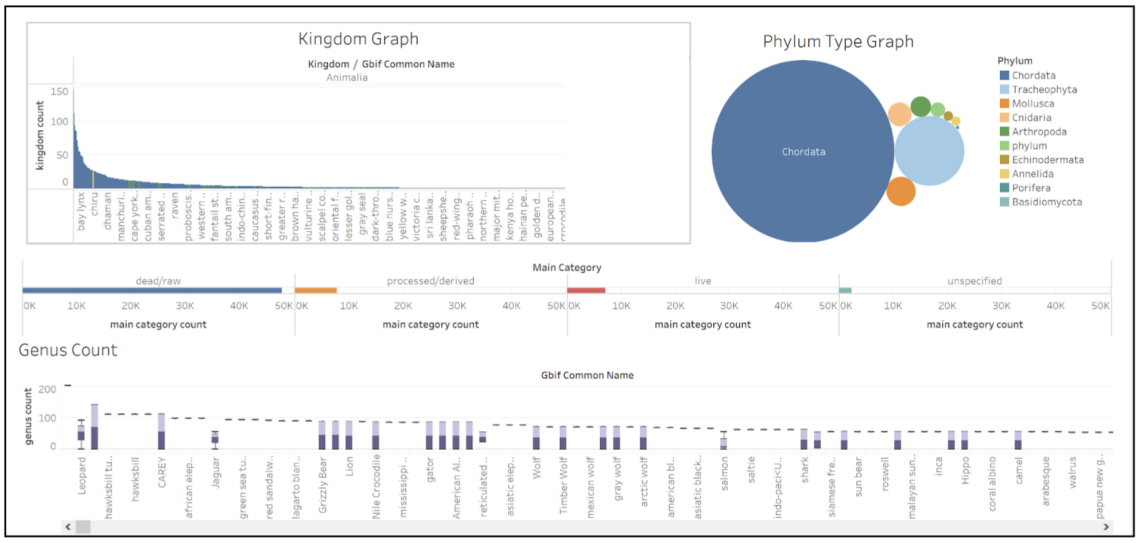


Fig. 3. Interactive Dashboard used to get insights on Kingdom-Family-Phylum-Genus based on various filters

dashboards, each dashboard portrays a story and pattern that we can observe. For Instance, In Fig. 3, you can observe that we've embedded information in various formats for the user to understand and visualise. By applying filter on various elements/graphs you can also observe the other graphs dynamically display results accordingly.

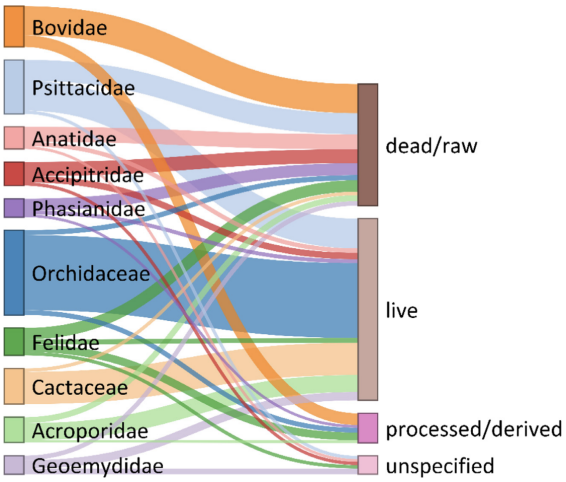


Fig. 4. Use-taxa combinations at the family taxonomic level

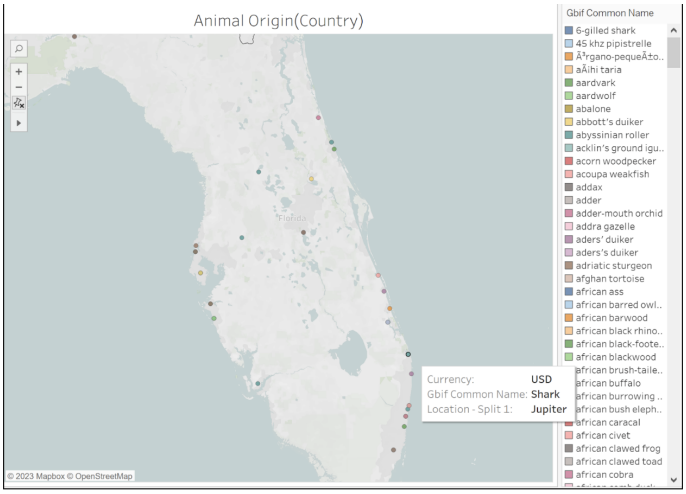


Fig. 5. We can see sharks and other kinds of fish are mostly traded from costal line of Florida, USA

### 5.1 Insights

We've made many graphs and derived insights based on them. Some of the results from our visualisations includes:

- dead (whole animal) has more records with Family: Accipitridae than others, and subcategory count tends to be higher for Accipitridae. This may explain why subcategory count is higher for dead (whole animal).
- The distribution of records for Kingdom is different for Wolf, dead animal than other marks, but no correlations were found with the measure values shown in the viz
- ebay.ph has more records with Gbif Common Name: Halibut than others, and domain count tends to be higher for Halibut. This may explain why domain count is higher for ebay.ph

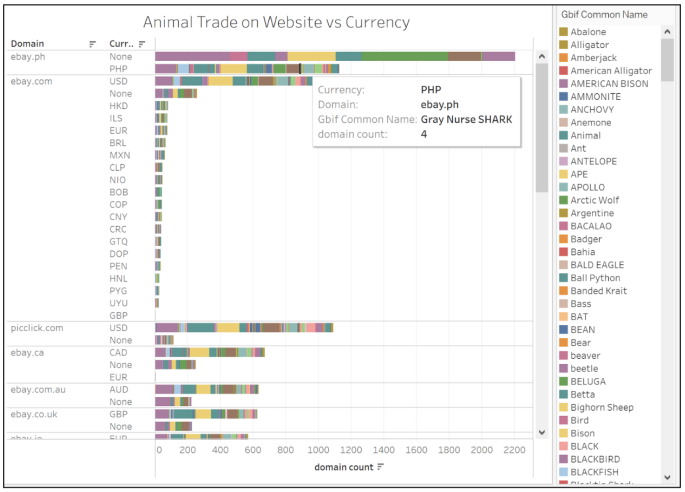


Fig. 6. One of many visualisations made, Animal Trade based on Currency vs Domain

- The live has more records with Genus: Amazona than others, and subcategory count tends to be higher for Amazona. This may explain why subcategory count is higher for live. live has more records with Family: Psittacidae than others, and subcategory count tends to be higher for Psittacidae. This may explain why subcategory count is higher for live.
- Ostrich is the bird that was traded the most in USD Currency. Followed by Sharks in USD which is related to location.
- leopard, Animalia, Chordata has more records with Genus: Panthera than others, and kingdom count tends to be higher for Panthera. This may explain why kingdom count is higher for leopard, Animalia, Chordata.
- dead animal has more records with Db Taxa Name: cheilinus undulatus than others, and standardized use type count tends to be higher for cheilinus undulatus. This may explain why standardized use type count is higher for dead animal.

## 6 ACKNOWLEDGMENTS

Thank you NYU Tandon and Prof. Juliana Freire for providing us with the datasets, dataproc resources, pipeline information and implementations. We really appreciate the help and guidance provided to us through out this project.

## 7 REFERENCES

[1] Oliver C. Stringham, Stephanie Moncayo, Eilish Thomas, Sarah Heinrich, Adam Toomes, Jacob Maher, Katherine G.W. Hill, Lewis Mitchell, Joshua V. Ross, Chris R. Shepherd, Phillip Cassey, Dataset of seized wildlife and their intended uses, Data in Brief, Volume 39,2021,107531,ISSN 2352-3409,https://doi.org/10.1016/j.dib.2021.107531.

[2] M. 't Sas-Rolfes, D.W.S. Challender, A. Hinsley, D. Veríssimo, E.J. Milner-Gulland Illegal wildlife trade: scale, processes, and governance Annu. Rev. Environ. Resour., 44 (2019), pp. 201-228, 10.1146/annurev-environ-101718-033253

[3] Juliana Barbosa, Sunandan Chakraborty, Juliana Freire, "A Flexible and Scalable Approach for Collecting Wildlife Advertisements on the Web" InfoWild '23, October 22, 2023. https://drive.google.com/file/d/1tCHJdNtie96Fkm9_NS-vUTZiq30s3wlJ/view.

[4] Justin Kurland & Stephen F. Pires (2017) Assessing U.S. Wildlife Trafficking Patterns: How Criminology and Conservation Science Can Guide Strategies to Reduce the Illegal Wildlife Trade, Deviant Behavior, 38:4, 375-391, DOI: 10.1080/01639625.2016.1197009.

[5] GBIF: The Global Biodiversity Information Facility, 2021. https://www.gbif.org/what- is- gbif

[6] P. Siriwat, V. Nijman Wildlife trade shifts from brick-and-mortar markets to virtual marketplaces: a case study of birds of prey trade in Thailand J. Asia-Pacific Biodiv. (2020), 10.1016/j.japb.2020.03.012

[7] O.C. Stringham, A. Toomes, A.M. Kanishka, L. Mitchell, S. Heinrich, J.V. Ross, P. Cassey, A guide to using the internet to monitor and quantify the wildlife trade, Conserv. Biol. (2021), doi: 10.1111/cobi.13675.

[8] S.A. Chamberlain, E. Szöcs, taxize: taxonomic search and retrieval in R, F10 0 0Res. 2 (2013) 191, doi: 10.12688/ f10 0 0research.2-191.v2 .

[9] The TRAFFIC International Wildlife Trade Portal, using their website (https://www.wildlifetradeportal.org/).

[10] CITES trade database, using their website (https://trade.cites.org/)