**Sourcing Open Data:**

**EMPRES Global Animal Disease Surveillance**

**Data Source/Summary:**

The [Global Animal Disease Surveillance Data](https://www.kaggle.com/datasets/tentotheminus9/empres-global-animal-disease-surveillance) (GADS) was sourced from Kaggle.com. Various hyperlinks allow for you to explore the [Food and Agriculture Organization of the United Nations](https://www.fao.org/animal-health/areas-of-work/early-warning-and-disease-intelligence/FAO's-EMPRES-Global-Animal-Disease-Information-System-(EMPRES-i-)/en) from which this data was collected. The data is external as FAO works closely with countries to obtain information and this data is available for public use according to their [Terms and Conditions](https://www.fao.org/contact-us/terms/en/).

This data contains information about disease outbreaks in wildlife and livestock animals from the last 8 years. It includes information about highly monitored diseases such as African swine fever, Foot and mouth disease and bird-flu. It organizes its information to show animals at risk, cases reported, deaths of animals and has mild reports of human exposure.

I’ve chosen this data set as it deals with animals and their disease prevalence. It’s an interesting topic to understand what disease can affect our wildlife and the livestock that provide us with food. The large scale of animals exposed and the number of those populations that either succumb to their disease or are slaughtered to protect other populations plays a great deal in how our wildlife and livestock cohabitate.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Data Profile:**

*Cleaning Data:*

Excel

* Capitalized column headers
* Ensured all data was aligned identically
* Renamed Columns as indicated below:
  + Admin1: City/Providence
  + observedDate: Obs.Date (D/M/Y)
  + reportingDate: Report Date (D/M/Y)
  + sumAtRisk: Sum Sp.AtRisk
  + sumCases: Sum Sp.Cases
  + sumDeaths: Sum Sp.Deaths
  + sumDestroyed: Sum Sp.Destroyed
  + sumSlaugthered: Sum Sp.Slaughtered
  + humansGenderDesc: Human Gender
  + humansAge: Human Age
  + humansAffected: Humans Affected
  + humansDeath: Human Deaths
* In Region column replaced all values titled Americas as North America (114 values)
* In Serotypes column all blanks replaced with Not Specified
* Blank values in the columns Sum Sp.AtRisk, Sum Sp.Cases, Sum Sp.Death, Sum Sp.Destroyed, Sum Sp.Slaughtered, Human Gender, Human Age, Humans Affected, Human Deaths are left blank in the excel spreadsheet.
  + *Excel Workbook for Data Profiling and Quality Measures will be submitted with report.*

Jupyter Notebook

* Renamed columns as indicated above
* Addressed Mix type data and corrected it as a string (str).
* Missing data for the following:
  + Sum Sp.AtRisk 7251
  + Sum Sp.Cases 2473
  + Sum Sp.Deaths 2840
  + Sum Sp.Destroyed 4003
  + Sum Sp.Slaughtered 4773
  + Humans Age 15940
  + Humans Affected 15591
  + Humans Deaths 16557
    - In 6.1 animal data imputed with ‘0’.
    - In 6.3 human data imputed with ‘0’.
    - *Gemini AI helped using an advanced code to change values to ‘0’.*
* In 6.3 Region value ‘America’ changed to ‘North America’ to match the JSON file obtained.
* *Gemini AI help develop an advanced code to create the Choropleth made with custom bins and a legend with more specific color ranges.*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Limitations & Ethics:**

According to Kaggle.com the limitations within this data include column description, source and the frequency with which the data is updated.

Exploring the data there is some uncertainty with the column descriptions and there is a lack of data completion with the columns concerning human exposure. Furthermore, the description on Kaggle.com states that the data is within 2 years. In fact the data holds information from no earlier than 2017 with the latest documentation being 1996. I couldn’t locate an earlier dataset from recent years.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Questions to Explore:**

1. What disease has the most cases?
2. What region has the most disease prevalence?
3. Is there a time of year that diseases are most likely to occur?
4. If data allows – how many humans die from the diseases surveyed?
5. Do animals commonly die from the disease they’re exposed to or are they slaughtered?
6. What disease has the highest death rate in animals?
7. Are wildlife more commonly affected or domesticated animals?