Data Immersion
Exercise 6.1
Jordan Novelli

Sourcing Open Data:

EMPRES Global Animal Disease Surveillance

Data Source/Summary:

The Global Animal Disease Surveillance Data (GADS) was sourced from Kaggle.com. Various hyperlinks allow for you to explore the Food and Agriculture Organization of the United Nations 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. 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.*
- There was an attempt to transpose the Sp.Description column in excel to get distinct counts of species for the disease. Excel couldn't transpose the Sp.Descrription column with the Disease column as it created more columns than Excel can handle. The original document is 17,009 rows and Excel can only handls 16,384 columns.

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.
- Exercise 6.5 – removed outliers to perform a more accurate kmeans cluster.
- *Gemini AI help develop an advanced code to create the Choropleth made with custom bins and a legend with more specific color ranges.*

_____

Additional Data Used:

- Custom GEOJSON file created using geojson-maps
- Time-Series Data taken from Quandl. This data is not related to the Global Animal Disease Surveillance data used in this project.

_____

Limitations & Ethics:

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

Data Immersion
Exercise 6.1
Jordan Novelli

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 in recent years. There were also limitations presented from the Sp.Description column. There are multiple species combinations in a single cell. Transposing this data wasn't available in excel.

_____

Questions to Explore:

1. What disease has the most cases?
   a. African Swine fever has the most cases with a count of 3,292 sum 8,833 in Europe.
2. What region has the most disease prevalence?
   a. Europe has the most disease prevalence with a % compared to Asia, Africa, and North America.
3. Is there a time of year that diseases are most likely to occur?
   a. Data limitations as the date is organized by MM/DD/YYYY.
4. If data allows – how many humans die from the diseases surveyed?
   a. With the small amount of human data available, I was able to determine that Asia had the most deaths with 153 deaths from the MERS-CoV, 140 deaths from Avian Influenza and 2 deaths from Rabies. These were null entries so I cannot determine age or gender of the humans that succumbed to death.
   b. Using that same data I can determine that an additional 53 males and 3 females succumbed from MERS-Cov and an additional 19 males and 5 females succumbed from Avian Influenza.
5. Do animals commonly die from the disease they're exposed to or are they slaughtered?
   a.
6. What disease has the highest death rate in animals?
   a. There are a variety of species that succumbed to a variety of disease:
      i. Domestic cattle were susceptible to as many as 10 different diseases including: Anthrax, Bluetongue, Bovine spongiform encephalopathy, Bovine tuberculosis, Brucellosis (Brucella abortus), Brucellosis (brucella suis), Contagious bovine pleuropneumonia, Foot and Mouth disease, Lumpy Skin disease, West Nile Fever.
      ii. Domestic goats were susceptible to as many as 9 different diseases including: Anthrax, Bluetongue, Brucellosis (Brucella abortus), Brucellosis (Brucella melitensis), Foot and Mouth disease, Peste des petits ruminants, Rabies, Rift Valley fever, Sheep Pox and Goat Pox.
      iii. Domestic goats and sheep combined were susceptible to as many as 7 diseases including: Anthrax, Bluetongue, Brucellosis (Brucella melitensis), Foot and Mouth disease, Peste des petits ruminants, Rift Valley fever, Sheep Pox and Goat Pox.
      iv. Domestic sheep were susceptible to as many as 6 diseases including: Anthrax, Bluetongue, Brucellosis (Brucella melitensis), Foot and Mouth disease, Peste des petits Ruminants, Sheep Pox and Goat Pox.

> > v. Avian Influenza affected a wider variety of animals from domestic avians, wild avians, and even some mammals. Domestic chickens, domestic unspecified birds, and domestic ducks were more likely to succumb to avian influenza.
> > vi. Wild boars were more likely to have African Swine Fever than domestic swine.
>
> b. The highest deaths were from Avian Influenza with 2,514,962 deaths and Newcastle Disease with 280,586 deaths. Then Africa Swine Fever, Peste des petits ruminants, Porcine reproductive and respiratory syndrome, Anthrax, Bluetongue and Lumpy Skin Dis had less than or equal to 9,336 deaths.

7. Are wildlife more commonly affected or domesticated animals?
   a. Captive/Domestic animals are more likely to die from disease than wildlife.