# Correlation Analysis: Zoonotic Diseases and Agricultural Trade patterns

*Authors: Blythe Weng, Tahreem Karim, Che-Yu Liu*

## Motivation:

We were all interested in doing a health related project, and decided to focus on zoonotic diseases.

*With Covid-19 still fresh on everyone's mind, we wanted to analyze zoonotic outbreak data from before 2019. That is how we decided to focus on the years 2015-2019.* Specifically, we wanted to know where the most documented cases are and what trade products those countries have in common.

To this end, we created **visualizations** to help characterize the data. Then we took our (merged) dataset containing production volumes and Zoonotic disease cases and ran a **regression** to see if there is any relationship between the most frequently produced items and Zoonotic cases. This project could be a small step towards helping policy experts and epidemiologists characterize future outbreaks.

## Background:

*A disease is zoonotic by definition if it is naturally transmitted between animals and humans.* Zoonotic diseases can spread between human and animal through body fluids, animal bites, eating infected meat, and contaminated water. List of commonly known zoonotic diseases include bird flu, west nile disease, anthrax, rabies, and many other commonly occuring diseases. See the CDC's website for more details (before it's gone!).

## Objectives:

The questions we want to answer are as follow:

1. Can we characterize Zoonotic disease outbreaks **before Covid (2019)**?
2. Are there any interesting spikes/trends in zoonotic cases within certain regions or countries?
3. Is there an association between the products these outbreak ridden countries produce and the volume of outbreaks?
4. And how confident are we in these results/connections we are trying to make?

# Data Sources

| Name | Description | Size | Access |
|---|---|---|---|
| WAHIS quantitative dashboard - World Animal Health Information System | Official reports submitted by veterinary services that include:<br>• A **variety of diseases**<br>• Where outbreaks occurred<br>• When they occurred<br>• Which species were involved<br>• Severity & case counts<br>• Whether a pathogen is endemic or introduced | 139,000 rows,<br>19 columns<br>(19.8 mb) | https://wahis.woah.org/#/dashboards/qd-dashboard<br>Format: **CSV** |
| FAOSTAT - Food and Agricultural Organization of the United Nations | A database of **crops and livestock products** as our secondary database. It includes comprehensive information on from around the world, including:<br>• Country production volumes<br>• Quantities of live animal trade<br>• meat, poultry, dairy, eggs, hides (bovine, swine, sheep)<br>• Yearly trade patterns<br>• Trading partners (import-export links) | 1705 rows,<br>15 columns<br>(0.25 mb) | https://www.fao.org/faostat/en/#data/QCL<br>Format: **CSV** |

# Methodology

## Data Manipulation Methods

We loaded and preprocessed our WAHIS data from a loaded **csv** that included data from 2015-2019. We only wanted data from these years because we didn't want COVID to skew our findings, and because it has already been studied endlessly. *We only kept rows that included zoonotic diseases.* Dropping those rows reduced the size of the dataset to something much more manageable.

Before merging our two datasets, we needed to manipulate and reshape our production dataset since it was lacking a singular year column, that the WAHIS data set had. Instead, it had multiple year columns (ex: y2018, y2019) with 0 or 1 values indicating whether the event happened in that specific year. We had to remove all these year columns and consolidate the year values within a singular, general year column. We also noticed that our cases column included non-numerical dash values. However *with our merge, we wanted to group by country, year, and disease and sum the outbreak case number together.* As a result, we had to clean out data by removing the dash values and turning them into NANs so we could sum them together. The reason we didn't turn case value into 0 is because it would mess with averages we would be taking for later visualizations.

We also noticed that there were some countries that had **naming mismatches** between the two dataframes, specifically between Russia and Turkey, with it being listed as Russian Federation and Türkiye (Rep. of) in the production database. As a result, we had to map theses two different values to each other and change the naming inconsistencies to just Turkey and Russia. *Another issue was that China was being listed as 4 separate regions within the WAHIS database* so we had to consolidate those as well. We consolidated the 4 regions: China, mainland, China (People's Rep. of), China, Hong Kong SAR, and China, Macao SAR all under China.

Lastly, to ensure our dataset was easy to manipulate, we removed any unnecessary columns we were not going to use, keeping only country, item, element, unit, year, world region, cases, country_std, & trade value.

# Methodology: Grouping and averaging

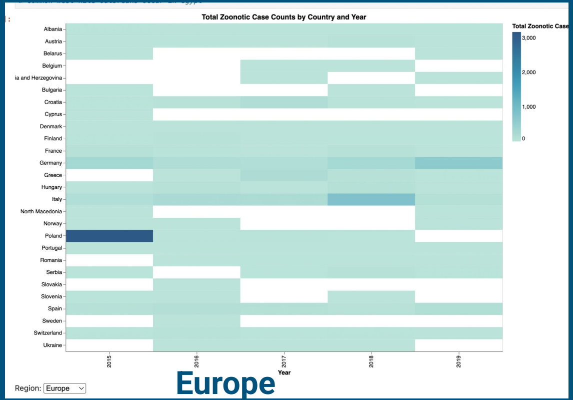
### Top 10 countries with most Zoonotic Cases found

For our visualization displaying the average Zoonotic cases for the top 10 countries, we had do some further grouping and manipulation. We started by grouping by country and year before calculating the average zoonotic cases found per country per year throughout the span of 2015-2019. After this grouping, we sorted by descending order to get the countries with the top 10 case values. For our visualization, we only decided to keep these 10 countries with the top values and display their average cases per year in a **lollipop visualization**.

### Correlation Matrix

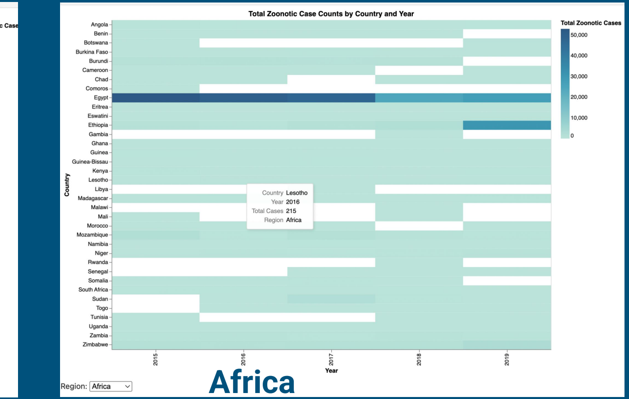
For our correlation matrix visualization, we **first** needed to identify the top 25 produced items within each country. This was done by using a groupby to sum the Trade_Value column across all years per country per item. **Next**, we wanted to categorize these *products into Low, Medium, and High frequency tiers to see if there was a correlation between frequently produced products and zoonotic cases.* **Frequency** was determined by counting how many countries each product appeared in as a top-25 item. The items were then sorted from most to least frequent and categorized using a tertile split. These tier classifications were merged back into merged_df2, where we then aggregated the data by country, year, and tier — summing trade values then pairing them with zoonotic case counts. **Finally**, this table was pivoted so that each tier became its own column variable, allowing us to compute a **Pearson correlation matrix** that served as the basis for both our correlation heatmap and subsequent regression analysis.

# Analysis: Heatmap of Zoonotic Case by Country and Year



Region: Europe

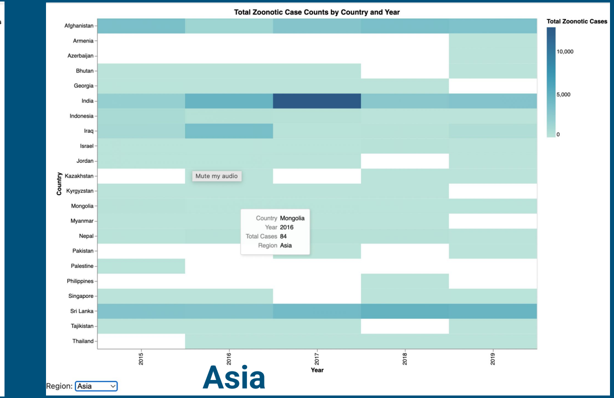
Europe



Country Lesotho
Year 2016
Total Cases 215
Region Africa

Region: Africa

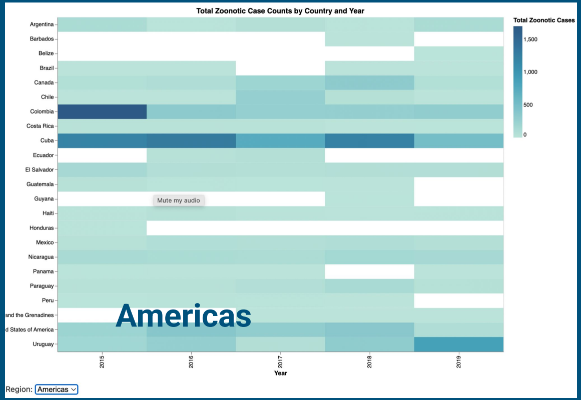
Africa



Country Mongolia
Year 2016
Total Cases 84
Region Asia

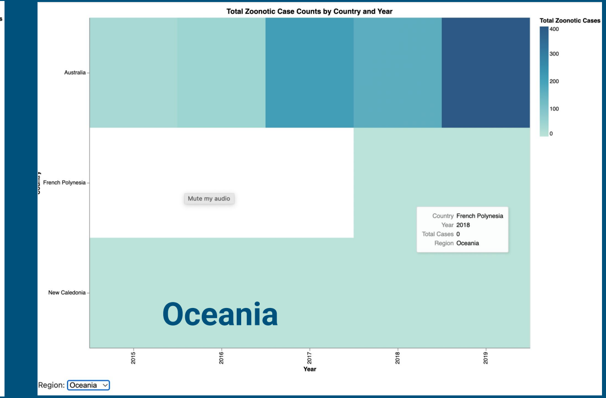
Region: Asia

Asia



Region: Americas

Americas



Country French Polynesia
Year 2018
Total Cases 0
Region Oceania

Region: Oceania

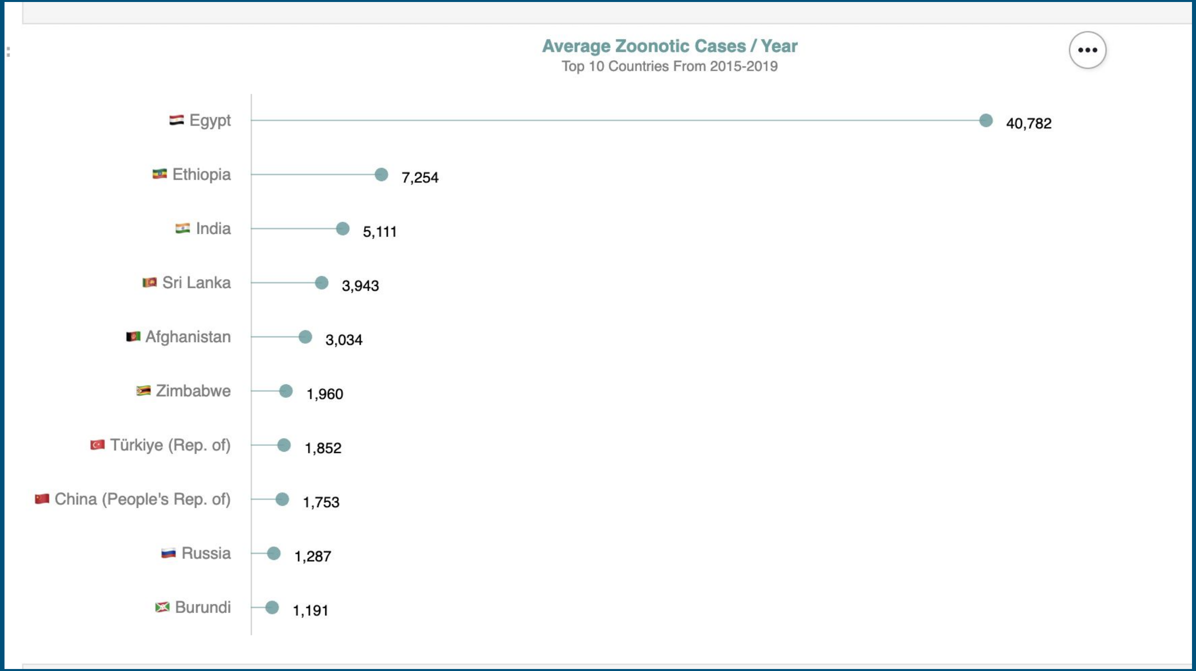
Oceania

**First** we evaluated the Zoonotic case count per country and year. For our visual, we decided to go with a heatmap because it would allow us to easily compare and visualize the distribution of cases between countries.

Including a selector for region greatly reduced the amount of countries displayed on screen allowing for the visualization to be less cluttered and easier to read.

Through these charts, we discovered some interesting information, for example: Poland in 2015 had a large number of zoonotic cases. A quick search indicated that *Poland indeed had an African Swine Fever outbreak in 2015 following the year it was first introduced in the country*

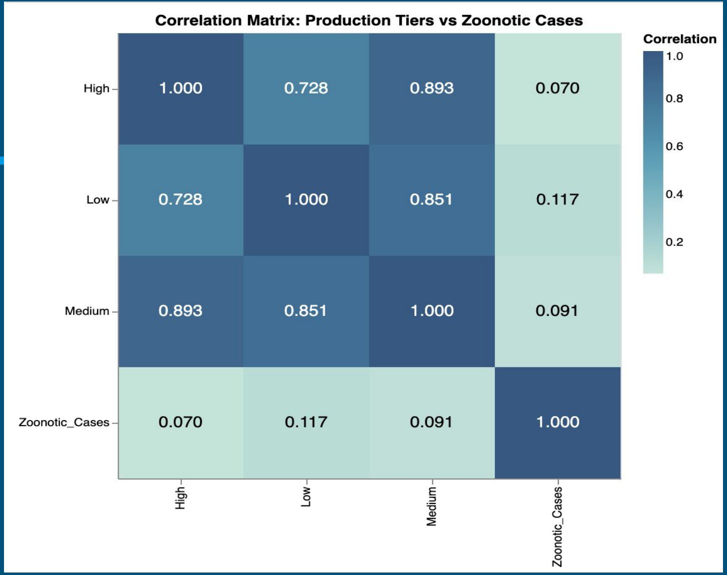
# Analysis: Lollipop Graph of Avg Zoonotic cases per year (Top 10 countries)



After seeing a general distribution of zoonotic cases per country and year, we were curious about which countries held the most average cases per year throughout the period of 2015-2019. We calculated the top 10 countries with the **highest amount of zoonotic cases** and created a horizontal lollipop graph for easy visual comparison.

Our results were very interesting, with *Egypt vastly leading the total amount of zoonotic cases,* while case averages for the other countries were relatively close. This could be due to Egypt's prominence of West Nile Disease *frequency tiers were determined based on separating out products by tertiles

# Analysis : Correlation Matrix & Regression Analysis


Correlation Matrix: Production Tiers vs Zoonotic Cases

|  | High | Low | Medium | Zoonotic_Cases |
|---|---|---|---|---|
| High | 1.000 | 0.728 | 0.893 | 0.070 |
| Low | 0.728 | 1.000 | 0.851 | 0.117 |
| Medium | 0.893 | 0.851 | 1.000 | 0.091 |
| Zoonotic_Cases | 0.070 | 0.117 | 0.091 | 1.000 |

*Figure 7.*The figure was obtained from a correlation matrix using Pearsons coefficient to see correlations between different frequency tiers* and zoonotic cases. This figure was then used as inspiration for creating a simple linear regression model to further confirm our findings.

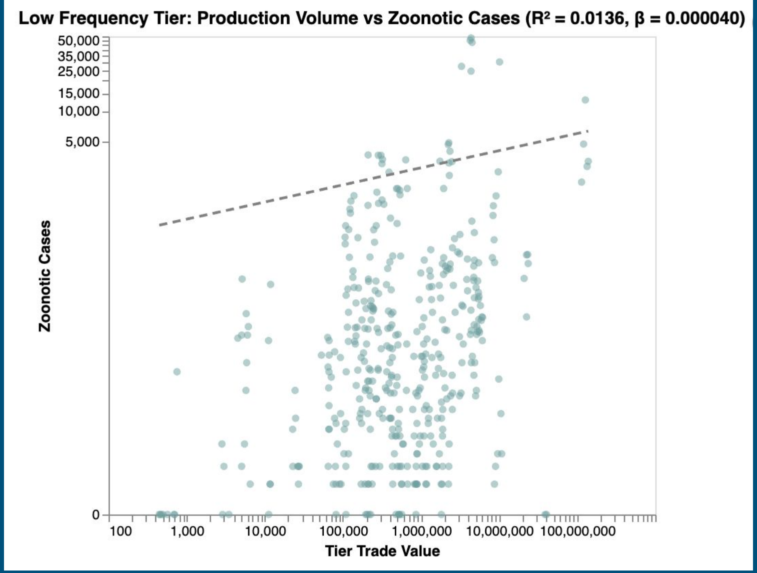

Low Frequency Tier: Production Volume vs Zoonotic Cases ($R^2 = 0.0136$, $\beta = 0.000040$)

*Figure 8.*The figure was obtained from the tier_agg dataframe, it displays the relationship between zoonotic case count and trade value amongst low frequency produced products. The line shows a slight positive correlation between the two values. However, due to the coefficient being 4.0e-5, we can see this is a very weak relationship

Now that we understand how zoonotic case counts are distributed across countries, *we want to take a closer look at the relationship between products frequently produced across all countries and their relationship with zoonotic case counts.* From the values obtained in *Figure 7* we can see that there exists a gradient amongst the production tiers. **Low frequency** products had the highest correlation (0.117) with Zoonotic Cases, followed by Medium frequency products and finally **High frequency** products. This figure was then used as inspiration to run a **Regression analysis** to determine how production volume equates to zoonotic case count. The regression had confirmed the pattern we saw in our correlation matrix where Low Frequency produced items had a higher positive correlation with zoonotic case count. However, with each respective R2 and Coefficient score obtained, *we determined that production value is a weak predictor of Zoonotic case counts.*

# Summary

1. **Can we characterize Zoonotic disease outbreaks before Covid (2019)?**

   Yes we are able to see the Zoonotic diseases were most prevalent in Africa and concentrated most in Egypt.

2. **Are there any interesting spikes/trends in zoonotic regions within certain regions or countries**

   Based on our heatmap results, we found some interesting data points such as an outbreak that occurred in Poland in 2015 or Cuba  leading Zoonotic cases found in the Americas. According to our top 10 most common average  zoonotic  cases lollipop graph, it seems that the Africa region had the most countries listed within that graph. Furthermore, Egypt far surpassed every other country in the top 10 which was also an interesting insight. Additionally, *though our study showed weak correlation, the least frequently produced items did have a distinctly higher correlation with Zoonotic cases.*

3. **Is there a relation between the products these outbreak ridden areas produce and the volume of cases?**

   Based on our regression analysis, *there is weak correlation between volume of outbreaks and products produced.*

4. **And how confident are we in these results/associations we are trying to make?**

   We are confident that there is little correlation between production value and zoonotic case count. As displayed from our correlation analysis and visualizations shown in an earlier slide.

## Next Steps/Further Continuations

Through our visualizations and analyses, we were able to obtain some **super interesting results**. For example noticing how Egypt was an outlier with a vast amount of zoonotic cases in comparison to any of the other countries. As a next step, it would be interesting to look further into trade routes within Egypt as well as other environmental factors that may play a role. Egypt has been known to be a hot spot for **West Nile Virus** so these further studies may reveal what environmental factors could be correlated with the spread of this disease.

It was also interesting looking through our heatmap and discovering certain countries such as Poland having a spike in Zoonotic diseases for just a singular year. This led to the uncovering of a lesser known outbreak, which could serve as a valuable resource for future epidemiologic simulations. At minimum, we realized that data is an unique way to rediscover history.

Lastly looking through our correlation matrix of production values versus zoonotic disease cases, it was interesting to see that there was no apparent correlation between production value and zoonotic case count. Although somewhat disappointing, it was still an illuminating and surprising conclusion. This leads to the question of whether other factors such as food handling or climate could have possible correlations, and is a testament to how difficult investigating pathogenesis of a zoonotic disease can be.

# Statement of Work - Main Contributors

## Jupyter ntbk 1

Merged Dataset

Two Visualizations

*Blythe & Tahreem*

**Mascot**

*Raja the baby Whippet*

## Jupyter ntbk 2

Regression Model

Two Visualizations

*Che-Yu*

*(ft. Tahreem &)*

## Report

Narrative

Slides

Proposal

*Blythe, Che-Yu, & Tahreem*

More technical and authorship details inside the GitHub Repository *(by Che-Yu!)*. For interactive visuals, the notebooks need to downloaded and re-uploaded to a local notebook.

Thank you!