**Exploratory Data Analysis Report on Lumpy Skin Disease**

Rana Hamed, Heba Ibrahim, Beshoy Ageeb, Eman Mansour

**Description**

This project focuses on analyzing Lumpy Skin Disease (LSD) using a dataset that includes various environmental and demographic factors, such as geographic coordinates, climate variables, and livestock populations. By employing statistical analysis and predictive modeling techniques, the project aims to uncover the relationships between these factors and the prevalence of LSD, providing valuable insights for disease management.

**Benefits of the Project**

* **Enhanced Disease Management**: Provides actionable insights into environmental factors influencing LSD outbreaks, enabling timely interventions.
* **Geospatial Analysis**: Identifies high-risk regions using geographic data, facilitating targeted health initiatives and monitoring.
* **Climate Impact Insights**: Explores how climate variables like temperature and precipitation affect disease dynamics, informing adaptive livestock management practices.
* **Predictive Modeling**: Develops models to forecast potential outbreaks, empowering stakeholders to take proactive measures to protect livestock.
* **Resource Optimization**: Guides effective allocation of veterinary services and vaccination programs based on disease prevalence and environmental conditions.
* **Temporal Trends**: Analyzes trends over time to improve forecasting of disease outbreaks and plan intervention strategies accordingly.
* **Public Awareness**: Increases understanding of Lumpy Skin Disease among farmers and stakeholders, promoting better practices in livestock management.

**Dataset** From

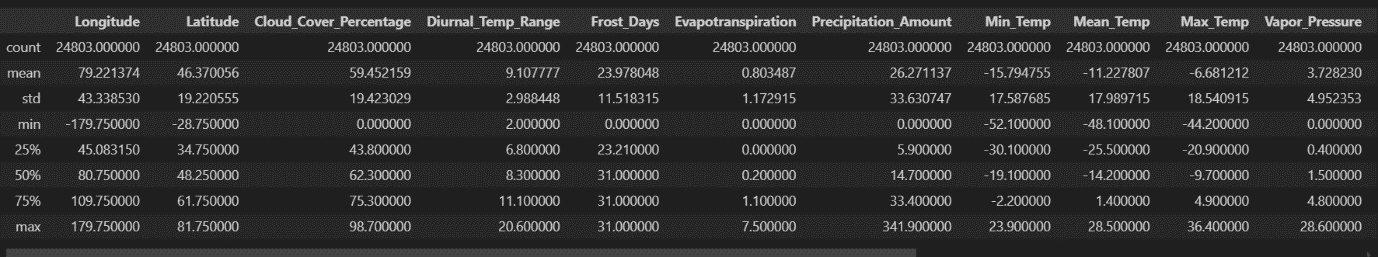
* **Source**: [Lumpy Skin Disease Dataset on Kaggle](https://www.kaggle.com/datasets/saurabhshahane/lumpy-skin-disease-dataset)

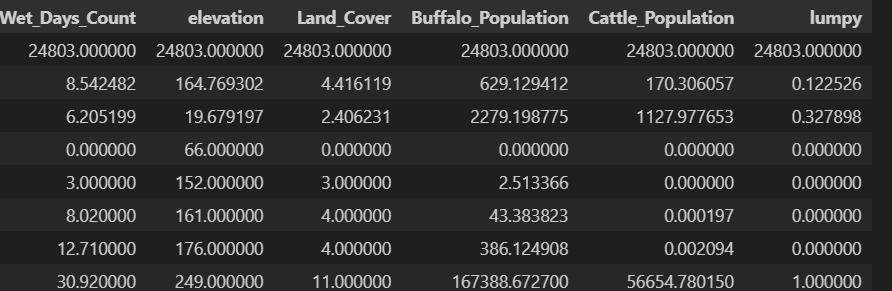
**Data Shape**: (24803, 20)

**Data** Info

1. **Longitude**: Geographic coordinate that specifies the east-west position of a point on the Earth's surface.
2. **Latitude**: Geographic coordinate that specifies the north-south position on the Earth's surface.
3. **Region**: Name or code for a larger geographical division within a country.
4. **Country**: Name or code of the country where the data point is located.
5. **Reporting Date**: Date when the data was recorded or reported.
6. **Cloud Cover Percentage**: Percentage of cloud cover.
7. **Diurnal Temp Range**: Difference between daily max and min temperatures.
8. **Frost Days**: Number of days with temperatures below freezing.
9. **Evapotranspiration**: Measures potential water evaporation and plant transpiration.
10. **Precipitation Amount**: Total rainfall in the area.
11. **Min Temp**: Daily lowest temperature.
12. **Mean Temp**: Average temperature in the region.
13. **Max Temp**: Daily highest temperature.
14. **Vapor Pressure**: Humidity levels, indicating moisture in the air.
15. **Wet Days Count**: Number of days with significant rainfall.
16. **Elevation**: Altitude of the location.
17. **Buffalo Population**: Density of buffalo per square kilometer.
18. **Cattle Population**: Density of cattle per square kilometer.
19. **Dominant Land Cover**: Type of land cover (e.g., forest, grassland, agriculture).
20. **Lumpy**: Result (0: Control, 1: Disease).

Data Description





**Data Cleaning Steps**

1. **Remove Duplicates**: Identified and removed 608 duplicate entries.
2. **Remove Unnecessary Columns**: Dropped columns (region, country, reportingDate) with over 90% missing values.
3. **Remove Unique Columns**: Excluded the column (dominant\_land\_cover) and removed rows where:
   * Latitude is not between -90 and 90.
   * **صورة تحتوي على نص, خريطة

     تم إنشاء الوصف تلقائياً**Longitude is not between -180 and 180.
4. **Box Plot and Histogram Analysis:** For (Cloud\_Cover\_Percentage, Diurnal\_Temp\_Range, Frost\_Days, Min\_Temp,Mean\_Temp, Max\_Temp, Wet\_Days\_Count ) columns.
5. **Remove outliers** **in the following columns**: Wet\_Days\_Count, Cloud\_Cover\_Percentage, Diurnal\_Temp\_Range, Frost\_Days.

**Data Visualization**

**Univariate Analysis**

**صورة تحتوي على نص, لقطة شاشة, رسم بياني, خط

تم إنشاء الوصف تلقائياً**

**صورة تحتوي على نص, لقطة شاشة, رسم بياني, خط

تم إنشاء الوصف تلقائياً**

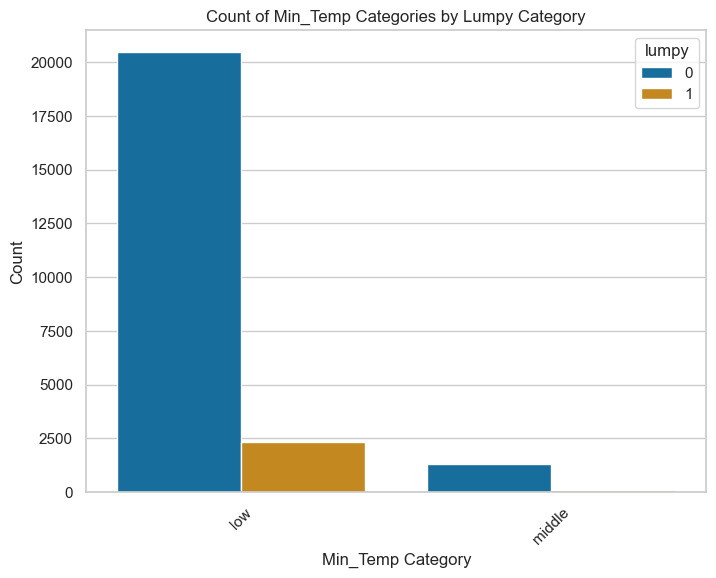
**صورة تحتوي على نص, لقطة شاشة, رقم, الخط

تم إنشاء الوصف تلقائياً**

**صورة تحتوي على نص, لقطة شاشة, رسم بياني, مستطيل

تم إنشاء الوصف تلقائياً**

**Bivariate Analysis**

**صورة تحتوي على نص, لقطة شاشة, مستطيل, رسم بياني

تم إنشاء الوصف تلقائياً**

**صورة تحتوي على نص, لقطة شاشة, رقم, تخطيط

تم إنشاء الوصف تلقائياًصورة تحتوي على نص, لقطة شاشة, رقم, تخطيط

تم إنشاء الوصف تلقائياً**

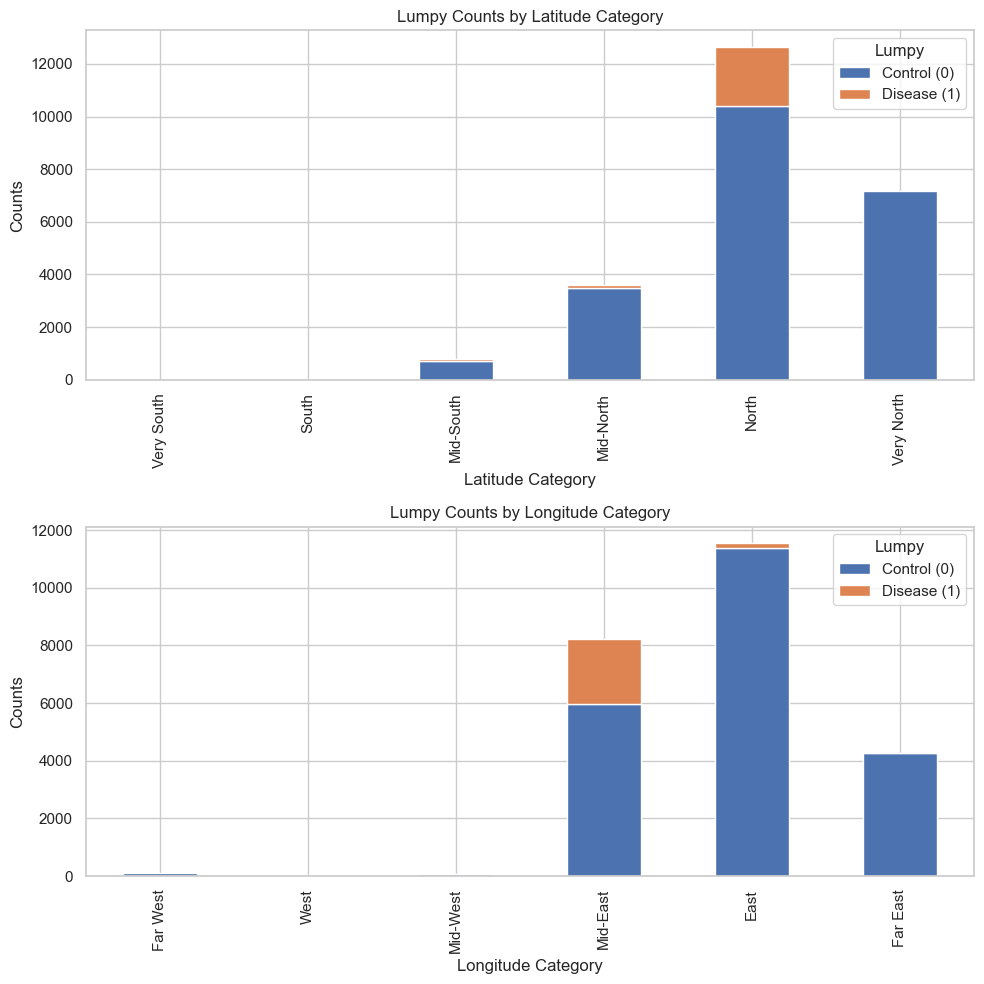
* **Converted the "Lumpy" column to categorica**l for easier analysis.
* **Converted the "Latitude" column and " Longtitude " to categorical using**:

lat\_bins = [-90, -60, -30, 0, 30, 60, 90]

lat\_labels = ['Very South', 'South', 'Mid-South', 'Mid-North', 'North', 'Very North']

long\_bins = [-180, -120, -60, 0, 60, 120, 180]

long\_labels = ['Far West', 'West', 'Mid-West', 'Mid-East', 'East', 'Far East']

****

**Map for size spread disease**

**صورة تحتوي على نص, خريطة, لقطة شاشة, رسم بياني

تم إنشاء الوصف تلقائياً**

**Preprocessing Steps**

1. **Correlation Analysis**: Identified high correlations among Min Temp, Max Temp, and Mean Temp; recommended removing one for simplicity.

**صورة تحتوي على نص, لقطة شاشة, شكل, ميدان/ مربع

تم إنشاء الوصف تلقائياً**

1. **Define Features and Target Variable**: Selected appropriate features for modeling.
2. **Split the Data**: Divided the dataset into training and testing sets.
3. **Handle Imbalanced Dataset**: Used RandomOverSampler to balance the classes.
4. **Scaling**: Applied StandardScaler for feature scaling.
5. **Feature Selection**: Used SelectKBest to select the top 10 features based on ANOVA F-statistic.