Longitudinal Data Analysis of Farm Animals

Surapot Nonpassopon
201731151

Supervised by Charles Taylor

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analitics
as part of the degree of

Master of Science in Data Science and Analystics

The University of Leeds, School of Mathematics

September 2024

The candidate confirms that the work submitted is his/her own and that appropriate
credit has been given where reference has been made to the work of others.

Abstract

This study employs longitudinal data analysis to investigate factors affecting the growth and weight of lambs and cows at Crag House Farm in Leeds, UK. By utilising Linear Mixed Models (LMMs), the research aims to achieve two primary objectives: identifying significant factors influencing animal growth and developing robust and accurate predictive models for livestock weight. The study finds that variables such as age, weaning weight, type, sire (father of animals) and weather conditions significantly impact the animals' growth. For example, older animals, higher birth weight, and specific sires contribute statistically significantly to greater weight gain for farm animals. The study also finds that the model after tune by utilising the cross-validation process with the feature selection technique demonstrates statistically significant high predictive accuracy compared to the model before tune ($p < 0.001$) in both terms of lambs and cows. The results from tuning the LMMs to predict the weight of lambs and cows are evaluated using root mean squared errors (RMSE), which are around 2.5 kg and 25 kg, respectively. These tuning models provide a valuable tool for farmers to predict animal weight. However, the study faces limitations related to data quality, computational resources, and adaptability to other farms, suggesting that future research should focus on improving these areas to enhance model accuracy, robustness, and adaptability across different datasets and species. Therefore, this research highlights the importance of data-driven approaches in farm management, offering a strong foundation for improving livestock growth and profitability.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Animal farming plays an important role in sustaining the global food supply and supporting rural economies. Understanding the factors that influence animal growth is essential for optimising farm management and ensuring sustainable agricultural practices. Therefore, research on farm animals is an important way to advance agricultural practices, ensure the well-being of animals, increase the health and growth of farm animals, and increase the quality of productivity on farms. By focusing on studying farm animals, the longitudinal study, which is the study that collects data from the same subjects over a long period (Arnold *et al.*, 2011; Jennings, 2005), could play a crucial role in this research. These studies enable researchers to observe changes and identify trends that might be missed in short-term studies (Arnold *et al.*, 2011). By tracking the same animals over time, longitudinal research could provide an overall view of how different factors influence growth over time (Jennings, 2005). Therefore, the longitudinal study offers deeper insights that could lead to more effective farm management practices.

This study utilises datasets from Crag House Farm in Leeds, UK, focusing on managing farm animals, especially lambs and cows. The dataset of farm animal records spans from 2016 to 2022 for lambs and from 2014 to 2024 for cows. Using data to gain insights into farm animal management is essential. It allows farmers to make informed decisions based on evidence from the analysis result rather than intuition, which sometimes might lead to the wrong decision. In terms of understanding the factors that influence the weight of farm animals, this is particularly important because weight is a crucial determinant of market value. Additionally, improving animal weight before sale can significantly increase profitability, making it a critical aspect of farm management.

Overall, this project aims to raise the strengths of longitudinal analysis to understand valuable insights from the recorded longitudinal data that could inform the farmer to improve the management practice and profitability of farm animals.

Figure 1.1: The Visit to Crag House Farm in Leeds to Inspect the Animals and Understand the Factors That Might Affect Them, as Discussed with the Farmers.

1.2 Objective

The main objective of this longitudinal farm animal analysis project is to comprehensively understand the factors that impact the growth and weight of lambs and cows on Crag House Farm in Leeds. The study aims to achieve two primary goals:

1. Identifying significant factors that affect the growth trends of these animals.

2. Developing the most accurate and reliable predictive model to assist farmers in forecasting the weight of their livestock.

The valuable insights from this research could provide farmers with useful, practical information, enabling them to make more informed decisions rather than relying solely on intuition when managing the farm. Consequently, farmers could develop robust management strategies that enhance the health and productivity of the animals while supporting sustainable growth and profitability of the farm.

1.3 Significance of the Study

Choosing the longitudinal study for farm animal analysis is important because it provides a detailed understanding of how various factors affect animal growth and development over time (Arnold *et al.*, 2011; Jennings, 2005). Longitudinal research collects data from the same subjects over an extended period, allowing researchers to observe long-term trends, unlike short-term studies (Liu, 2015; Murphy *et al.*, 2022). Therefore, this approach is particularly crucial in the context of farm animal management because the longitudinal studies could be used to identify the long-term effects of variables such as genetics, nutrition and weather conditions on the growth and health of farm animals. These insights are crucial for adapting animal management practices, which could lead to improved animal welfare and productivity. By understanding how these factors influence the growth of the animals, farmers could make decisions based on the evidence that enhances the efficiency and sustainability of their farm operations.

In terms of this study, this report's results could potentially impact farm management practices by providing valuable insights to farmers to create a management plan to improve livestock health and productivity. To elaborate, farmers in Crag House Farm might make more appropriate decisions about animal management by using the information findings from the longitudinal analysis, such as selecting the breeder that might provide the higher weight trend of offspring that could make the newborn calves trend to reach the market weight easily. This advantage of using the decisions based on information found through longitudinal analysis could lead to improved growth rates of the animals and better-estimated market weights to increase the profitability of the farm. Additionally, understanding the long-term effects of various factors on livestock can help farmers develop more sustainable farming practices and plans to enhance their operations' overall efficiency and sustainability.

1.4 Outline of Dissertation

This dissertation is organised into six chapters. This chapter is the introduction to the project. This is followed by Chapter 2, which introduces the methodology used in this study, including the approaches and techniques employed to address the research objectives. Chapter 3 describes the dataset utilised in the research and presents the process of data preparation, including steps such as data cleaning and transformation. In Chapter 4, the initial analysis is presented, including exploratory data analysis (EDA) and statistical analyses aimed at gaining insights from the dataset. Chapter 5 focuses on the longitudinal analysis of lambs and cows, presenting the analysis and results which archive the two main objectives of the study. Finally, Chapter 6 concludes the dissertation with a discussion of the results, summarising the key findings and offering recommendations for future research.

Chapter 2

Methodology

2.1 Introduction

This chapter outlines the research methodology that employs longitudinal design to analyse the quantitative and qualitative data, utilising linear mixed models (LMMs) and feature selection techniques for tuning the LMMs. These techniques are used to achieve the two goals of this research objective, which include understanding the factors that affect the weight of farm animals and developing accurate and reliable predictive models for animal weight.

2.2 Overall Research Design

This study employs a longitudinal study to analyse the quantitative data that collects and analyses numerical data to understand relationships between variables and explain observed outcomes using statistical methods (Ding *et al.*, 2018; Olanrewaju *et al.*, 2020). Moreover, qualitative data is also analysed in this study, allowing the study to explore patterns and relationships among categorical data, such as animal type, sex and birth type, therefore providing a more comprehensive understanding of the factors influencing farm animal growth (Bailey, 2008; Liamputtong, 2009).

In terms of the methodology covering the project objective, The primary focus of this research is on fitting a linear mixed model (LMM), which can handle the hierarchical data structure of farm animals (Murphy *et al.*, 2022; Pusponegoro *et al.*, 2017), to understand the factors affecting their growth rates. Following this, the LMMs are tuned by using input selection techniques to enhance predictive accuracy, identify model inputs that need careful recording to improve model accuracy and provide practical advice to farmers on key variables to monitor. An alignment of research design covers the research objective, which includes an explanation of the three main approaches and reasons for selecting the methodology, which will be provided in Section 2.2.1 and 2.2.2.

2.2.1 Fitting the Linear Mixed Model

- **Objective:** To identify and understand the factors that influence the growth rate of farm animals.

- **Approach:** The LMMs are used to fit the data and interpret the result received from the coefficient in the model.

- **Rationale:** The LMMs could be well-suited for this analysis as it handles repeated measurements and nested data structures, which are common structures found in longitudinal data (Murphy *et al.*, 2022). In terms of this study, the data of Crag House Farm are structured by repeated measurements of the same animal ID on different dates, recording the weight; therefore, the choice of LMMs could be suited for this study. Moreover, after fitting the model, LMM coefficients could provide a detailed understanding of how different factors impact growth over time (Murphy *et al.*, 2022; Pusponegoro *et al.*, 2017).

2.2.2 Tuning the Linear Mixed Model

- **Objective:** To enhance the robustness and predictive accuracy of the LMMs for animal weight and find the best accuracy of LMMs.

- **Approach(Steps):**

  1. **Cross-Validation:** Implement cross-validation to estimate the model's prediction accuracy by repeatedly dividing the data into training and validation sets to test and record the model performance in each iteration (Maleki *et al.*, 2022; Ranganathan *et al.*, 2019). This method could be used to reduce the problem of overfitting, which is when the model performs well on the training data only while poorly performing on unseen datasets (Subramanian and Simon, 2013; Ying, 2019).
  2. **Feature Selection:** Use feature selection techniques to determine the most relevant variables for predicting animal weight.
  3. **Evaluate Akaike Information Criterion (AIC):** Record AIC during an iteration of the cross-validation process to compare the goodness of fit of statistical models while penalising for model complexity (Toga, 2015). Compare the average AIC across models to identify the top five models with the lowest average AIC, suggesting high performance.
  4. **Evaluate Root Mean Squared Error (RMSE):** Use the top five models with the lowest average AIC to perform 100 rounds of cross-validation and record the RMSE, which measures the square root of the average squared differences between the actual and predicted values. This provides an error measure in the same units as the original data (Neill and Hashemi, 2018; Schneider and Xhafa, 2022).

5. **Identify Best Model:** Select the model with the lowest recorded RMSE that suggests robust and high-accuracy results.
6. **Compared the Before and After Tune Performance:** Compared the RMSE before the before and after tune of the LMMs to understand whether the model performance improves or not.
7. **Interpret the Tuned LMMs:** Compare the coefficients of the LMMs before and after tuning to understand how the tuning process has affected the model. This comparison will help identify which factors have become more or less significant in predicting lamb weight after tuning, allowing us to determine which variables play a crucial role in improving lambs' weight and which do not.
8. **Guide Farmers:** Use the best-performing model's input variables to advise farmers to record the key input factors for high prediction accuracy. Applying this model in real-world farm settings helps farmers predict animal weight more effectively, supporting improved animal management practices.

- **Rationale:** Using cross-validation and feature selection to tune LMM models ensures the inclusion of the most relevant variables. This approach improves the model's accuracy and robustness in predicting the animal weight (Maleki *et al.*, 2022). Moreover, Farmers can improve data collection by identifying and carefully recording these key input variables from the best accuracy model, leading to better-informed decisions and enhanced farm operations.

2.3 Explanation of Analytical Methods

Four main analytic methods that are used in this study will be described in this part.

2.3.1 Linear Mixed Model (LMM)

The Linear Mixed Model (LMM) is an extension of the linear model used to analyze data with hierarchical structures in which observations are nested within higher-level groups (Liu, 2015; Murphy *et al.*, 2022; Pusponegoro *et al.*, 2017). For example, measurements taken from multiple lambs(subjects) over time (repeated measures) create a two-level hierarchy data structure, as shown in Figure. The figure represents the hierarchy structure data of Lambs in Crag House Farm, with two levels of data by **Level 1:** Represents the individual lambs (for example, Lambs ID731, Lambs ID732 and Lambs ID733) and **Level 2:** represents the repeated measurements for each lamb at different ages (e.g., Age 3 months, Age 4 months).

LMMs could be useful for fitting longitudinal data, where repeated measurements are taken on the same subjects over time (Liu, 2015). LMMs could account for both fixed effects, which are consistent across all observations, and random effects, which vary across groups or subjects (Liu, 2015; Rao *et al.*, 2011). To elaborate, the term β represents the coefficient in LMMs and is used to quantify the magnitude and direction of the effects that various factors have on the output variable, which in this context is animal weight. Additionally, the term u_i in Equation 2.1 is used to calculate the random effect that occurs from differences among animals (subjects). This term captures the variability between individual animals, allowing the model to account for the unique characteristics of each animal that are not covered by the fixed effects (Murphy *et al.*, 2022; Rao *et al.*, 2011).

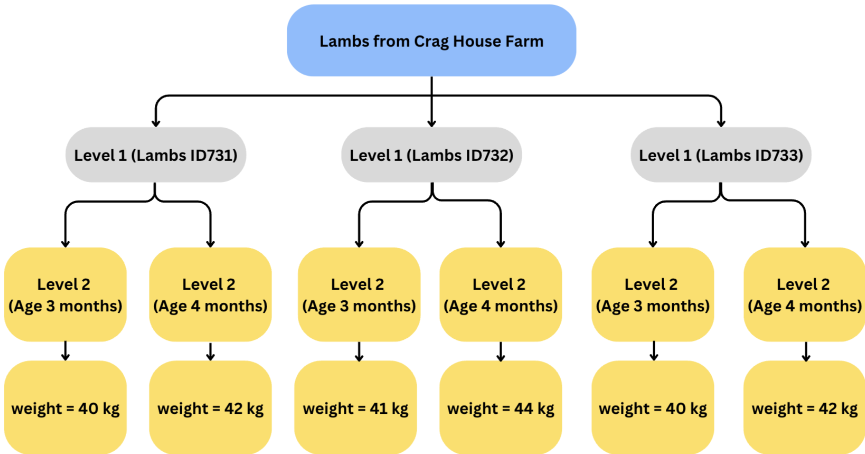

Figure 2.1: Hierarchical Structure of Longitudinal Data for Lambs' Weight Measurements at Different Ages: Grouping by Individual Lamb IDs from Crag House Farm with Repeated Observations of Weight

The equation of LMM is demonstrated in Equation 2.1 by including some explanation of the parameters.

The general form of the LMM is:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + u_i + \epsilon_{ij} \tag{2.1}$$

where:

- y_{ij}: Response variable for observation j in group i.
- β_0: Overall intercept.
- $\beta_1, \beta_2, \ldots, \beta_p$: Fixed effect coefficients.
- $x_{ij1}, x_{ij2}, \ldots, x_{ijp}$: Fixed effect predictors.
- u_i: Random effect for group i, representing variations between groups. It is assumed to follow a normal distribution with mean 0 and variance σ_u^2. This term accounts for the unobserved variability among the groups.

- ϵ_{ij}: Residual error term for observation j in group i, representing the within-group variability. It is assumed to follow a normal distribution with mean 0 and variance σ_ϵ^2.

2.3.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a standard evaluation metric used to measure the square root of the average of the squared errors, which are the differences between the actual values and the predicted values, as shown in Equation 2.2 (Neill and Hashemi, 2018). RMSE is particularly useful in farm animal analysis because it assesses the accuracy of a model by providing an error measure in the same units as the predicted values (as kilograms in this study). This metric is often used to compare the accuracy of different models through a cross-validation process, where lower RMSE values indicate better model performance, as the predicted values are closer to the actual values (Neill and Hashemi, 2018; Schneider and Xhafa, 2022).

The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{ij})^2} \tag{2.2}$$

where:

- n is the number of observations.
- y_i is the actual value.
- y_{ij} is the predicted value from LMM.

2.3.3 Cross-validation (CV)

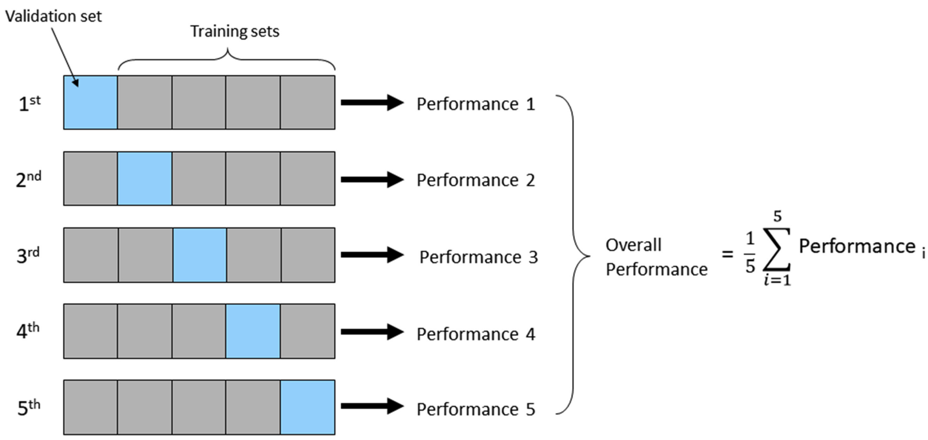

Figure 2.2: The Example of the Cross-Validation Process by Splitting the Data Set Into Five Folds to Use the Validation Sets to Test the Model to Receive the Performance of the Five Models, Moreover, the Overall Performance of the Model Is Determined by Averaging These Performance Values (Jumah *et al.*, 2022).

Cross-validation (CV) is a statistical method used to evaluate the predictive performance of a model by partitioning the original dataset into a training set to train the model and a validation set to test the model (Maleki *et al.*, 2022; Ranganathan *et al.*, 2019). For example, Figure 2.2 represents the cross-validation process by which each data fold is rotated as the validation set while the rest are used for training the model across multiple iterations. The advantages of this process are preventing the overfitting of the model and providing robust performance estimates by averaging results across iterations (Jumah *et al.*, 2022; Ranganathan *et al.*, 2019).

2.3.4 The Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a refined method that evaluates a model's ability to predict future values based on its fit to the current data (Mohammed *et al.*, 2015). It is used to compare the goodness of fit of statistical models while penalising for the complexity of the model as terms k in the model below (Toga, 2015). The formula for the Akaike Information Criterion (AIC) is provided as Equation 2.3:

$$\text{AIC} = 2k - 2\ln(L) \tag{2.3}$$

where:

- k is the number of parameters in the model.
- L is the likelihood of the model.

The AIC value will be increased if the number of input parameters the k increases. That might not be good because Lower AIC values indicate better models, as they balance fit quality with model simplicity by penalising the increase of additional parameters (Leiter and Turner, 2017).

2.4 The Overall Steps of the Project

This study has been divided into five main steps, as shown in Figure 2.3. The details of each will be provided in each subheading and information below.

Step 1: Data Understanding

Collecting and reviewing the datasets of farm animals from Crag House Farm and the dataset of weather conditions from Leeds City Council by inputting the data into the R to understand their structure, variables, and content.

Step 2: Data Preparation

This step involves joining the dataset of lambs by combining records from multiple years into a longitudinal data form or nested structure. After that, the longitudinal structure datasets are used to select valuable columns or factors for analysis, such as sex and type of animals. Additionally, necessary operations are performed, such as finding other useful factors to analyse by subtracting dates of birth from the weighing date to determine the age of animals at the weighing date and joining datasets of farm animals with weather conditions to analyse the impact of weather on animal growth. All these processes will be explained in detail in the data preparation section (Section 3.3) in Chapter 4.

Step 3: Exploratory Data Analysis (EDA): This step involves conducting initial analyses through visualisation techniques, such as line plots, to observe the growth trends of farm animals across different sexes. Additionally, simple statistical observations will be generated to identify patterns and insights in the data. These visual and statistical analyses could help to understand the distribution, trends, and relationships within the data. Moreover, the EDA might provide a foundation for more detailed analysis in subsequent steps.

Step 4: Data Analysis:

This phase is divided into two steps to capture the project's two main objectives.

1. Fit the LMMs and interpret the results to understand the factors affecting farm animal growth.

2. Tuning the LMMs through cross-validation, incorporating feature selection techniques to enhance model accuracy and reliability.

Step 5: Summary and Discussion:

In this phase, the discussion will link the results and insights of this study with existing literature, report the project's limitations, and suggest future directions for further research.

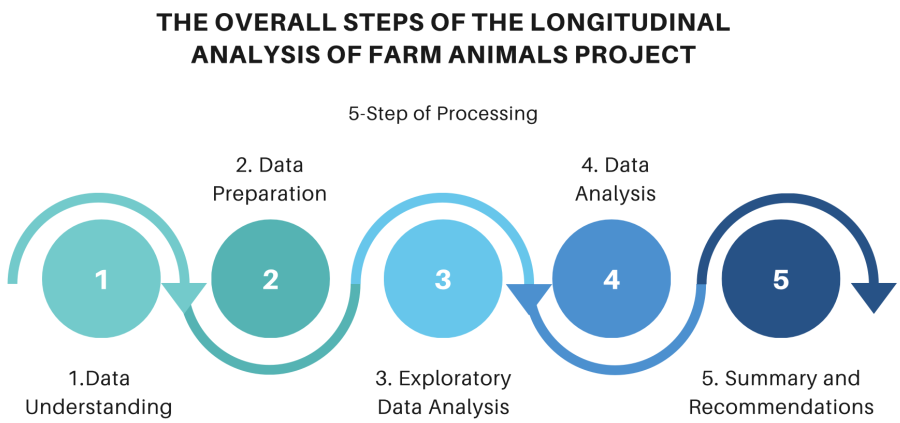

Figure 2.3: The overall steps of the longitudinal analysis study of the farm animal research

Chapter 3

Data Description

3.1 Data Source

The data for this study was collected from Calg House Farms, the local farm in Leeds, UK, focusing on records of newborn lambs from 2016 to 2022 and cattle from 2014 to 2024. The two datasets include detailed information about each lamb and cattle, recorded from birth until various stages of their growth and development. To elaborate, after the lambs and cattle are weaned, measurements are taken, and the weight on the other stage is recorded for the following month. Moreover, essential variables, such as sex or type, have been recorded in the dataset, which could allow for the analysis of other factors that affect animal weight. Additionally, each animal's exact date of birth is documented, and a timeline for subsequent measurements is provided, which could be used to calculate and track the age of animals.

In addition to the farm animal records from Crag House Farm, this study also utilises meteorological data published by Leeds City Council. This dataset, recorded hourly from 2000 to 2019 through the weather station at Leeds (Longitude: -1.543237, Latitude: 53.787769), includes crucial weather-related factors that may impact the growth of farm animals. For example, the key features are average wind speed, temperature, global radiation, and humidity (Leeds City Council, 2020). By incorporating these weather variables into the longitudinal analysis, this analysis could provide a comprehensive understanding of how environmental conditions influence the growth and weight of lambs and calves over time.

3.2 Data Explanation

The datasets used in this project will contain three main datasets: lamb datasets from 2016 to 2022, cow datasets from 2014 to 2024, and the weather condition dataset from 2000 to 2019. All three datasets will explain the variables in Table 3.1, 3.2 and 3.3.

3.2.1 Lamb Datasets Variables Explanation

Variables	Explanation
LambID	A unique identity number for each lamb.
Sex	The sex of the lambs (M for male, F for female).
D.O.B.	Date of birth of the lamb.
Dam	Identifier for the lamb's mother.
Sire	Identifier for the lamb's father.
Weaning Weight	The weight of lamb on the weaning day.
Dead Weight	Dead weight of the lamb (Carcass weight after slaughter).
Type	Type of lamb which is divided into two main types in this dataset which are: 1.Pedigree lambs are purebred lambs bred for breeding purposes to maintain breed standards. 2.Commercial lambs are crossbred lambs whose primary purpose is for meat production.
Non-Fixed Date Weight	Records the weight of lambs on non-fixed dates after weaning, represented as pairs of (date, weight in kg).

Table 3.1: Explanation of Variables in the Lamb Datasets

3.2.2 Cow Datasets Variables Explanation

Variables	Explanation
Calf no.	A unique identity number for each calf.
Sex	The sex of the calves (M for male, F for female).
Calf weight	The weight of cows on the weaning day.
Date Born	The date of birth of the calf.
Dam no.	Identifier for the calf's mother.
Sire	The name of the calf's father.
Non-Fixed Date Weight	Records the weight of cows on non-fixed dates after weaning, represented as pairs of (date, weight in kg).

Table 3.2: Explanation of Variables in the Cow Datasets

3.2.3 Weather Dataset Variable Explanation

Variables	Explanation
Date	The date of the recorded data.
Wind Speed	The daily average wind speed (meters per second).
Temperature	The daily average temperature in degrees Celsius.
Global Radiation	The daily average of global radiation in watts per square meter, representing the intensity of sunlight.
Humidity	The daily average of relative humidity in percentage, indicating the moisture content in the air.

Table 3.3: Explanation of Weather Dataset

3.3 The Process of Preparing the Animal Datasets

The step of preparing the lamb's and cow's datasets will be provided below, including the example of the final combined lamb and cow datasets after preparing in Figure 3.1 and 3.2, respectively, which could contain all key variable factors that are appropriate for the process of the part of the analysis.

1. **Combine Data:** Integrate the data of farm animals born in different years datasets into a combined dataset.

2. **Data Transformation:** Perform transformations on the data to obtain valuable factors for longitudinal analysis, including:

   - Transfer the data into a longitudinal structure, as shown in Figure 2.1, by aligning each animal ID with its corresponding weight recorded on different dates. This allows the dataset to include multiple entries for the same animal ID, each associated with a different date of weight measurement.
   - Calculate the age of the animal on the weighing day using the formula $\frac{\text{specific weighing day - D.O.B.}}{365.25}$ to convert the duration into years.
   - Identify the season during which the weight was recorded using the Met Office's definitions: spring (March, April, May), summer (June, July, August), autumn (September, October, November), and winter (December, January, February) (Met Office, 2024).
   - Specify the type of birth (single, twin, triplet, quadruplet) by matching lambs born on the same date and with the same mother to count the number and classify them as single, twin, triplet, or quadruplet births. In terms of this study, it could be found that different types of birth only occur in the lambs dataset, while cow datasets contain only a single birth.

3. **Joining Animal Dataset with Weather Dataset:** Merge the lamb and cow data with the weather condition data based on the corresponding dates. However, there might be some limitations on the cow dataset due to the majority of cow recording weights between growth periods being recorded after the year 2020, while the weather condition dataset covers only the years between 2000 and 2019. Therefore, this longitudinal analysis of cow datasets will not cover the effect of the weather conditions, which will be explained more in the section on limitations observed from datasets.

4. **Obtained Combined Dataset:** In terms of the combined lamb dataset, after processing the three steps above, the final combined lamb dataset used for the main analysis in this study includes data on lambs born in different years and weather conditions, which represent the example of the dataset in Figure 3.1. Key variables for analysis in this study

include lamb ID, Age of lambs, weight in kg, weaning weight, sex, type, sire (specific father of lambs), age of mothers (years), birth types (single, twin, triplet, or quadruplet), season of weight recording, average wind speed, temperature, global radiation, and humidity on the weight recording day.

ID	Age	Weight	WeaningWeight	Sex	Type	SireID	AgeOfMother	BirthType	WeightSeason	WindSpeed	Temperature	GlobalRadiation	Humidity
738	0.003	42	42	M	pedigree	1621	2.992	Twin	spring	9.433	7.987	160.384	73.681
738	0.294	42	42	M	pedigree	1621	2.992	Twin	summer	2.821	12.754	84.560	88.971
738	0.369	47	42	M	pedigree	1621	2.992	Twin	summer	2.703	23.685	390.592	68.307
738	0.444	50	42	M	pedigree	1621	2.992	Twin	summer	3.931	15.682	298.186	71.330
738	0.544	58	42	M	pedigree	1621	2.992	Twin	autumn	2.478	14.870	26.028	97.112
739	0.003	36	36	F	pedigree	1621	2.992	Twin	spring	9.433	7.987	160.384	73.681
739	0.294	36	36	F	pedigree	1621	2.992	Twin	summer	2.821	12.754	84.560	88.971
739	0.369	46	36	F	pedigree	1621	2.992	Twin	summer	2.703	23.685	390.592	68.307
739	0.444	47	36	F	pedigree	1621	2.992	Twin	summer	3.931	15.682	298.186	71.330
739	0.544	52	36	F	pedigree	1621	2.992	Twin	autumn	2.478	14.870	26.028	97.112
739	0.628	58	36	F	pedigree	1621	2.992	Twin	autumn	1.696	9.170	54.637	91.438
739	0.719	57	36	F	pedigree	1621	2.992	Twin	autumn	2.188	9.203	12.940	98.483
739	0.814	55	36	F	pedigree	1621	2.992	Twin	winter	3.491	8.271	103.014	81.187

Figure 3.1: Combined Lamb Dataset Used in Longitudinal Analysis Captured From R Which Selected Variables Including ID of Lambs, Age in Years, Weight (kg), Weaning Weight (kg), Sex, Type, Sire ID, Age of Mothers (Years), Birth Type, Season of Weight Recording, Wind Speed (m/s), Temperature (°C), Global Radiation (W/m²), and Humidity (%)

Regarding the combined cow dataset, after processing the three steps above, the combined cow dataset includes data on cows born in different years and weather conditions (in NA values), which represents the example of the dataset in Figure 3.2. The main variables for longitudinal analysis in this study include calves' specific ID, the age of cows, the weight of calves, the weight of calves on the weaning date, sex, the specific sire name or ID, the season the weight was recorded, and the age of the breeder (mother). As mentioned before, the weather conditions that are not covered on the weight day are represented as missing values (NA) in Figure 3.2.

CalfID	Age	Weight	CalfWeight	Sex	Sire	AgeOfMother	WeightAtSeason	WindSpeed	Temperature	GlobalRadiation	Humidity
700291	0.003	34	34	F	Nitro	4.077	Winter	NA	NA	NA	NA
700291	0.736	220	34	F	Nitro	4.813	Autumn	NA	NA	NA	NA
700291	0.851	242	34	F	Nitro	4.927	Winter	NA	NA	NA	NA
700291	0.947	285	34	F	Nitro	5.023	Winter	NA	NA	NA	NA
700291	1.024	321	34	F	Nitro	5.100	Spring	NA	NA	NA	NA
700291	1.131	339	34	F	Nitro	5.207	Spring	NA	NA	NA	NA
700291	1.788	447	34	F	Nitro	5.863	Winter	NA	NA	NA	NA
700291	1.864	461	34	F	Nitro	5.940	Winter	NA	NA	NA	NA
700291	1.922	465	34	F	Nitro	5.997	Winter	NA	NA	NA	NA
700291	2.113	474	34	F	Nitro	6.189	Spring	NA	NA	NA	NA
700291	2.190	518	34	F	Nitro	6.265	Spring	NA	NA	NA	NA

Figure 3.2: Combined Cow Dataset Used in Longitudinal Analysis Captured From R with Selected Variables Including ID of Calves, Age in Years, Weight of Calves, Weight of Calves on the Weaning Date, Sex, Type, Sire ID, Birth Type, Season of Weight Recording (While Unmatched Weather Conditions Are Represented as Missing Values - NA)

3.4 Limitation Observed from the Data

The problems with data quality and some limitations that need to be addressed will be explained in these parts. Firstly, even after the preparation processes, some values are still missing in the dataset, with 27% of data missing in the combined lamb dataset and 30% of data missing in the combined cow dataset. These problems are likely due to a lack of a continuous process for recording the data. It is important to inform farmers about the necessity of improving the standard of data recording to enhance data quality. This improvement will be crucial in the longitudinal analysis process, potentially providing more accurate results if continuous and complete data records are maintained. To ensure the quality of the analysis, rows of data containing missing values will be removed from the dataset before proceeding with the analysis.

Secondly, in the process of joining the combined cow dataset (which includes data on cows born in different years) with the weather dataset, some limitations were identified. The weather data, measured from 2000 to 2019, was found to be incompatible with the cow growth measured with the data measured after the weaning day, which started recording observations for cows born after 2020 (starting record in 2020 until now). Therefore, the growth weight data recorded for cows throughout their growth period does not match the weather condition dataset. As a result, this combined data cannot be used to analyse the effect of weather on cow growth. This problem could be solved by finding reliable weather condition data from Leeds City Council that provides the weather conditions recorded from the weather stations in Leeds (Leeds City Council, 2020). By updating the weather condition data to match the cow growth data dates, the analysis could become meaningful from the insight into the weather conditions that affect the farm animals.

Chapter 4

Initial Data Analysis

4.1 Descriptive Statistics

Some statistical results and insight on the lamb, cow, and weather datasets, which could be useful to report to the farmer to develop the farming management strategy, will be explained and summarised in this section.

4.1.1 Descriptive Statistics of Lamb Datasets

Number of Lambs Categorised by Sex, Type, and Birth Type

Variables	2016	2017	2018	2019	2020	2021	2022
Sex							
Male Lambs	41	40	27	35	49	41	38
Female Lambs	58	23	53	51	44	42	36
Type							
Pedigree Lambs	99	15	80	69	66	62	53
Commercial Lambs	0	48	0	17	27	21	21
Birth Type							
Single Lambs	12	17	11	16	15	11	18
Twin Lambs	52	36	56	52	54	46	44
Triplet Lambs	27	6	9	18	24	18	12
Quadruplet Lambs	8	4	4	0	0	8	0
Total Lambs born in each year (lambs)	99	63	80	86	93	83	74

Table 4.1: Comprehensive Analysis of Annual Lamb Births Categorised by Sex (Male and Female), Type (Pedigree and Commercial), and Birth Type (Single, Twin, Triplet, Quadruplet) from 2016 to 2022

Table 4.1 provides a comprehensive look at the number of lamb births by highlighting the variation of lambs by sex, type, and birth type over the 2016 to 2022 period. Some main insights that could be reviewed from Table 4.1 are as following paragraphs.

In terms of sex distribution, in most years, the number of female lambs born is slightly higher than that of male lambs. However, in 2017, the number of female lambs was significantly lower than that of male lambs. This anomaly might occur due to different breeding practices, for example, different selection of fathers or mothers of lambs and some changes in genetic linkage.

As regards the type distribution of lambs, pedigree lambs are consistently higher in number than commercial lambs each year. However, the number of pedigree lambs dropped significantly in 2017 to 15, while the number of commercial lambs increased to 48 in the same year. This drop could be linked to the fewer female lambs that same year. Because female lambs in the Crag House Farm are often kept for breeding to maintain or increase pedigree lines, a decrease in female lambs in 2017 could directly impact the number of pedigree lambs.

Regarding birth type, twin births are the most common type of birth each year, with numbers ranging between 36 and 56 lambs per year. Single and triplet births also occur regularly but are lower in number compared to twin births. Quadruplet births are quite rare, occurring only in a few years, with the highest number being 8 in both 2016 and 2021.

Statistical Analysis on Weaning Weight

The initial analysis of the weaning weight of the lambs contains the two main box plots provided in Figure 4.1, which will be provided in detail in the following paragraph.

In terms of Figure 4.1(a), this figure represents the fluctuations in weaning weights between male and female lambs over the years. This fluctuation indicates that the difference in weaning weights between sexes varies from 2016 to 2022. To understand whether the weaning weights of male and female lambs are statistically different or not across all years, this could be done to the independent t-test, which is used to compare the similarity of the average value in different groups of subjects (De Canditiis, 2019; Thukral *et al.*, 2023). In this context, an independent t-test is used to test the hypothesis that the weaning weight of lambs in different sexes across all the years are equal.

Setting hypothesis as:

- H_0: The average weaning weight of male lambs and female lambs across all years are equal.

$$H_0 : \mu_{\text{male}} = \mu_{\text{female}}$$

- H_1: The average weaning weight of male lambs and female lambs across all years are not equal.

$$H_1 : \mu_{\text{male}} \neq \mu_{\text{female}}$$

After comparison, the average weaning weight of male lambs is around 40 kg, which is higher than female lambs, which stands at 36 kg. The 95% confidence interval from the t-test by comparing the weaning weight across all years of male to female lambs, in which the lower and upper boundaries are -6 and -3 kg, rejects the null hypothesis (H_0) that the average weaning weight of lambs in different sexes is equal. And suggest that the male lambs have a higher weaning weight than female lambs, around 3 to 6 kg. The p-values, lower than 0.001, indicate the statistically significant difference in weaning weights between male and female lambs (De Canditiis, 2019; Loftus, 2022). Therefore, the result from the statistical test could be used to support the hypothesis that the weaning weight of male lambs is higher than that of female lambs.

Figure 4.1: Box Plots of Lamb Weaning Weight by Sex and Birth Type from 2016 to 2020 which includes: (a) Box Plot of Weaning Weight by Sex Across Different Years and (b) Box Plot of Weaning Weight by Birth Type Across Different Years

Regarding Figure 4.1(b), the figure illustrates that the weaning weight also varies based on the lambs' birth type, with lambs born as singles generally having higher weaning weights compared to twins, triplets, and quadruplets. To support this statement, the individual t-test could be used to compare the weaning weight of lambs by different birth types, as shown in Figure 4.2.

Figure 4.2 shows that single-born lambs have the highest median weaning weights, followed by twins, triplets, and quadruplets. The weaning weight decreases as the birth type order increases, indicating that single-born lambs are generally heavier at weaning than those born as multiples. The statistical analysis could support this assumption by the independent t-test, in which the result suggests highly statistically significant differences ($p < 0.001$) between the weaning weights of single-born lambs and all other birth types. There is also a statistically significant difference ($p < 0.01$) between twins and triplets, with twins heavier. No statistically significant differences were found between the weaning weights of single, twin, and triplets, which were paired with quadruplets. This might be because of the limitation of the low data on quadruplets-born lambs. Overall, the figure could support the assumption that single-born lambs tend to have higher weaning weights than twin triplet or quadruplet lambs. This insight could help the farmers to enhance feeding programs during the weaning period to support lamb growth and modify breeding strategies to produce heavier lambs.

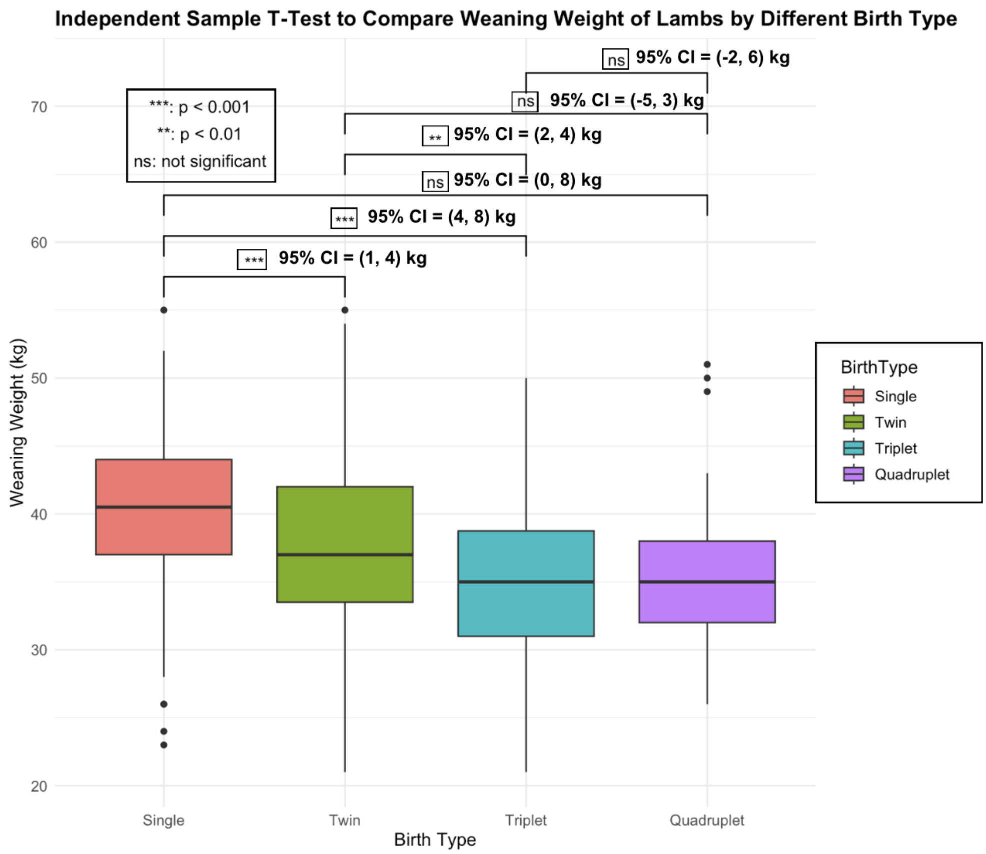

Figure 4.2: Independent Sample T-Test Comparing Weaning Weights of Lambs Across Different Birth Types Using Data from All Years: Single, Twin, Triplet, and Quadruplet

4.1.2 Descriptive Statistics of Cow Datasets

Number of Cows Categorised by Sex

The number of calves born each year by different sexes can be obtained from Table 4.2, which shows the slightly decreasing total numbers between 2014 and 2016. Moreover, the trend has some fluctuating birth numbers between 2016 and 2024, with values ranging from 12 calves to 19 calves. While most years represent a balanced distribution of male and female calves, some years show notable imbalances. For instance, 2019 had a higher number of male calves at 12 calves compared to female calves at 5 calves.

Birth year	The number of Calf's born in each year by different sex (calves)		
	Male calves	**Female calves**	**Total**
2014	6	11	17
2015	8	8	16
2016	6	6	12
2017	9	6	15
2018	7	8	15
2019	12	5	17
2020	9	8	17
2021	7	9	16
2022	5	9	14
2023	10	9	19
2024	4	13	17

Table 4.2: Descriptive Statistics of Calves Births by Sex (Male and Female) Over a Decade from 2014 to 2024

Statistical Analysis on Weaning Weight

The box plot, as Figure 4.3, illustrates some critical insights that weaning weights show noticeable variability from 2014 to 2024, with no consistent upward or downward trend. Male calves seem to have higher median weaning weights than female calves, with high noticeable differences in some years, such as 2014 and 2019. Certain years, such as 2017 and 2020, show narrow interquartile ranges of weaning weights, indicating less variability, while other years, such as 2015 and 2018, display broader interquartile ranges, suggesting more significant variability. The data includes several outliers, particularly in 2018, in which one individual male calf was significantly heavy at more than 57 kg, which was higher compared to other years, which might need to be rechecked before analysis that this is the actual value or mistake from keying or record the data because the outliers could be effect to the result of the analysis. Overall, the figure suggests that male calves have higher weaning weights than the female calves.

The analysis of the weaning weight of cows can be observed in Figure 4.3. Regarding Figure 4.3(a), the box plot illustrates some critical insights that weaning weights show noticeable variability from 2014 to 2024, with no consistent upward or downward trend. Male calves seem to have higher median weaning weights than female calves, with high noticeable differences in some years, such as 2014 and 2018. This insight could be used to make the assumption that male calves might have higher weaning weights than females.

To support the assumption, the independent t-test needs to be progressed to compare the weaning weight between the male and female calves to test the hypothesis that the average weaning weight of female and male calves is equal by using the data from all years recorded. And setting the null hypothesis as: $H_0 : \mu_{\text{female}} = \mu_{\text{male}}$. After performing the statistical test, the result provided the mean weaning weight of females, which was around 38 kg, while the males were approximately 41 kg. The 95% confidence interval for the difference in weaning weight between female and male calves is (-5, -2) kg with a p-value lower than 0.001, suggesting rejection of the null hypothesis that the weaning weight of female and male calves is equal. This implies that male calves might have a higher weaning weight than female calves by approximately 2 to 5 kg. Thus, the result from the t-test supports the assumption that male calves have higher weaning weights than female calves.

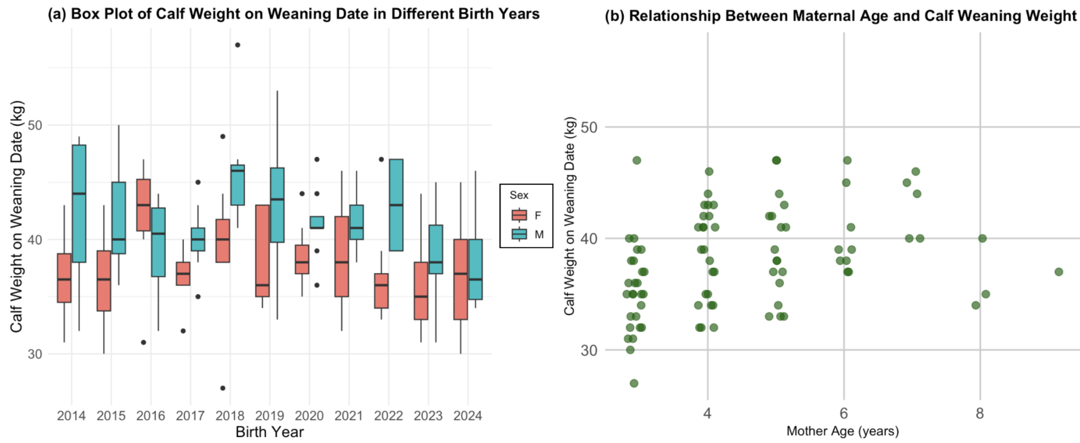

Figure 4.3: Analysis of Calf Weaning Weight includes (a) Box Plot of Calf Weaning Weight Distribution by Sex (Male and Female) Across Different Birth Years from 2014 to 2024 and (b) Scatter Plot of the Relationship Between Maternal Age and Calf Weaning Weight

In terms of Figure 4.3(b), the scatter plot demonstrates no clear linear relationship between maternal age and calf weaning weight. Calf weaning weights seem consistent across different maternal ages, suggesting that maternal age may not significantly impact calf weaning weight.

4.1.3 Descriptive Statistics of Weather Dataset

The statistical table of weather conditions in Leeds from Table 4.3 demonstrates some of the statistical values of four main factors: wind speed, temperature, global radiation, and humidity, which might affect farm animal growth. A summary of the table and how weather factors could affect animal growth will be provided in the following paragraph.

In terms of wind speed, the data shows the mean is significantly higher than the median at around 23 m/s, suggesting the presence of extreme outliers that skew the average, indicating a highly variable wind speed distribution, which distribution could be observed in Section 4.2.2 on Figure 4.4. The effect of animal growth from wind speed is that high wind speeds could cause stress and energy loss, which might reduce the growth rate of the animals (Saraux *et al.*, 2016).

As regards temperature, it ranges from around -4 °C to 26°C, with an average of around 11 °C. The effect of animal growth from temperature: If it is too hot or cold, it could cause cold and heat stress that might affect the growth rate of animals.

Regarding global radiation, this factor varies from 0 to 493 W/m², indicating a variety of solar exposure levels. This might affect animal growth due to global radiation, which could affect grass quality, which is essential for lamb nutrition (Trnka *et al.*, 2007).

In terms of humidity levels, the value ranges from 48% to 99%, with an average of 77%, showing relatively high humidity of weather overall in Leeds, UK. This factor might have some impact on the farm animals; with too high humidity could increase the risk of diseases and heat stress, while too low humidity can affect grass and hydration (Ford and Thorne, 1974).

Overall, weather conditions can significantly affect the growth of farm animals. Therefore, including weather effects in the longitudinal analysis would be beneficial in covering the effect of weather conditions on animal growth.

Statistical value	**Wind speed (meters per second)**	**Temperature (degree Celsius)**	**Global radiation (watts per square meter)**	**Humidity (percentage)**
Minimum	0.54	-3.93	0.00	48.38
1st Quartile (Q1)	2.62	6.86	73.59	70.37
Median	3.67	10.71	151.37	77.27
Mean	27.44	10.78	169.76	77.35
3rd Quartile (Q3)	5.46	14.83	247.50	84.61
Maximum	321.12	26.24	492.57	99.02

Table 4.3: Descriptive Statistics of Weather Dataset Collected from Weather Station at Leeds (Longitude: -1.543237, Latitude: 53.787769) from 2000 to 2020 (Leeds City Council, 2020)

4.2 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) of the farm animal data will examine various aspects through visualisation plots. Histograms will be used to observe the distribution of the data. Line plots will be employed to examine the growth trends of the animals. A correlation matrix will be utilised to identify correlations within the numerical data of farm animals. Significant insights from the EDA will be presented in the following section, divided into the different EDAs for lambs and cows.

4.2.1 Exploratory Data Analysis of Lamb Datasets

Distribution in Lamb Datasets

Figure 4.4 illustrates histograms showing the distribution of various numerical columns in the combined lamb dataset, which combined lamb datasets from 2016 to 2022 together with the weather datasets. In terms of the animal factors, the dead weight, weaning weight and weight of lambs show a normal distribution, with most values clustering around the mean. The age of lambs has a right-skewed distribution, indicating most lambs that were recorded as younger. The mother's age when giving birth to the new lambs is concentrated around a few specific values, indicating that the mother's common breeding ages are around 2 to 3 years old.

Regarding weather conditions, wind speed is highly right-skewed, with most values being low. This wind speed insight could be used to answer the question of why the average wind speed value is too far from the median value (around 27 and 4 m/s, respectively) in Section 4.1.3 in Table 4.3 from the histogram provides a lot of outlier values. Meanwhile, temperature and global radiation histograms display a more uniform distribution with some variation. The humidity histogram shows a normal distribution centred around 75 %.

Overall, the valuable insights gained from the EDA of histograms highlight the need for individualised strategies to improve the management practices of the farm. For example, the skewed distribution of wind speed and the variability in temperature and global radiation suggest the necessity of investing in protective structures for animals, such as shelters or shade, to mitigate the effects of extreme weather on lamb health and growth.

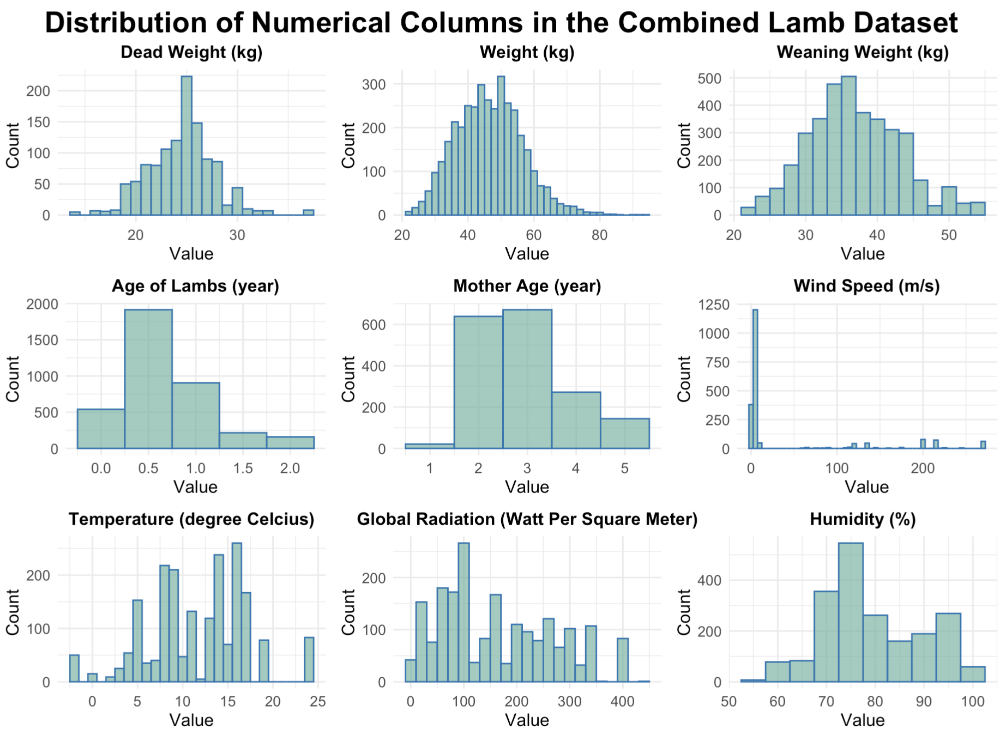

Figure 4.4: Histograms Showing the Distribution of Numerical Columns in the Combined Lamb Dataset, Including Animal Factors (Dead Weight, Weight, Age of Lambs, Duration Before Lambs Were Slaughtered, Age of Mother) and Weather Conditions (Wind Speed, Temperature, Global Radiation, Humidity)

Example of Growth Trends of Lambs

In terms of Figure 4.5(a), it reveals that both male and female lambs demonstrate steady weight growth over time. However, male lambs with the blue colour generally show higher weight gain compared to female lambs with the red colour, with male lambs reaching higher weights by the end of the observation period at around 430 days (approximately one year two months periods). Regarding Figure 4.5(b), commercial lambs (green) tend to have higher weight growth trajectories than pedigree lambs (brown) over the same period.

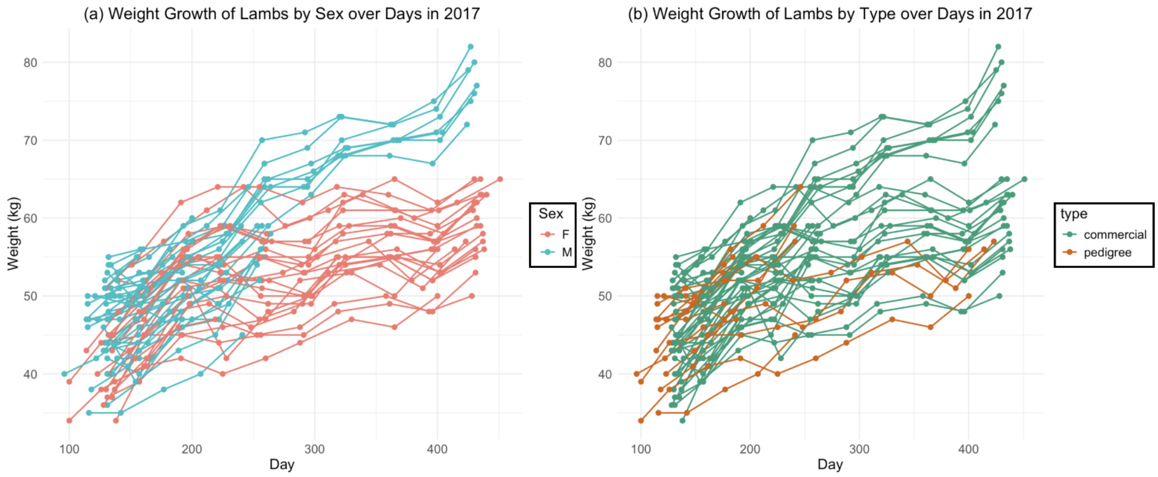

Figure 4.5: Weight Growth Trajectories of Lambs Over Time in 2017: (a) Categorised by Sex (red represents females and blue represents males) and (b) Categorised by Type (green represents commercials and brown represents pedigrees)

Correlation of the Combined Lamb datasets

Some main insights observed from the correlation matrix of the combined lamb dataset in Figure 4.6 include the following: the lamb weight is positively associated with both the age of the lamb (0.64) and weaning weight (0.50), indicating that older and heavier lambs at weaning generally have higher overall weights this could make as the assumption to analysis through the LMMs. Additionally, there is also a moderate positive correlation between dead weight and weight (0.42), indicating that lambs with higher live weights tend to have higher dead weights. Moreover, the mother's age seems to have a positive correlation with the weight and weaning weight of the lambs. This insight could suggest that the higher age of the breeder might give the son a higher weight. Furthermore, a strong positive correlation between global radiation and temperature (0.71) shows that higher temperatures are associated with higher global radiation levels. In contrast, global radiation and humidity exhibit a strong inverse relationship (-0.87), with higher radiation correlating with lower humidity. Several factors, like weight and temperature, show weak or no significant correlations at 0, suggesting their influence of temperature on lamb growth may have no effect.

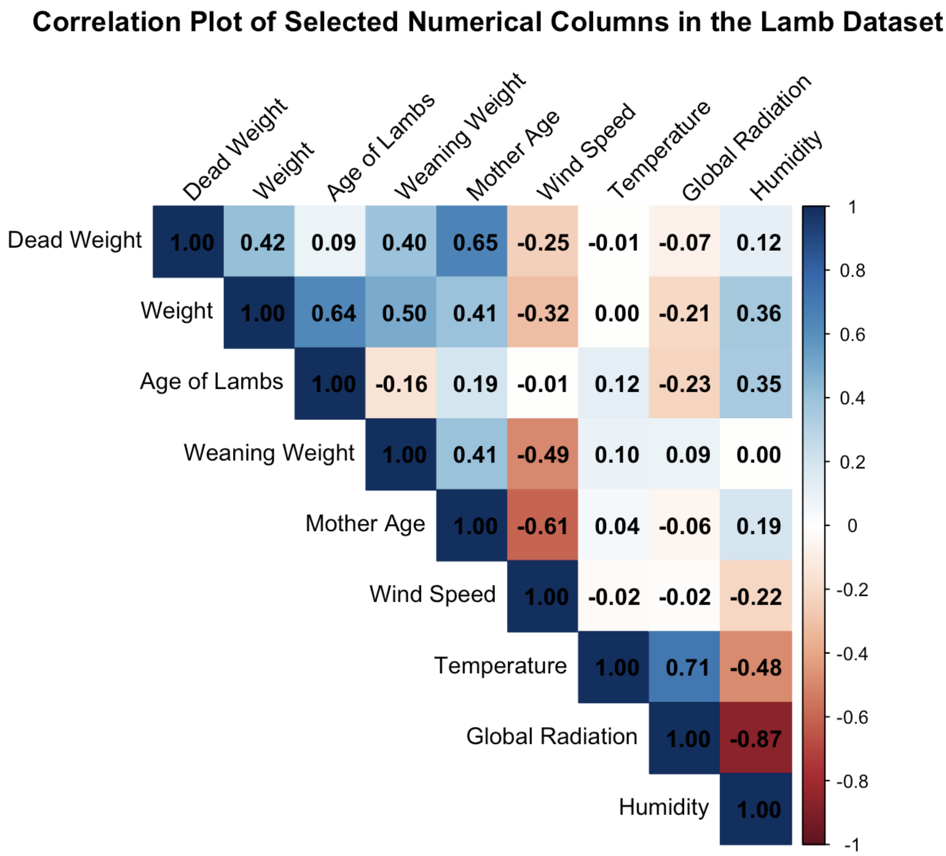

Figure 4.6: Correlation Matrix of Numerical Columns in the Combined Lamb Dataset (Blue indicates a positive relationship, while red indicates a negative relationship)

4.2.2 Exploratory Data Analysis of Cow Datasets

Distribution in Cow Datasets

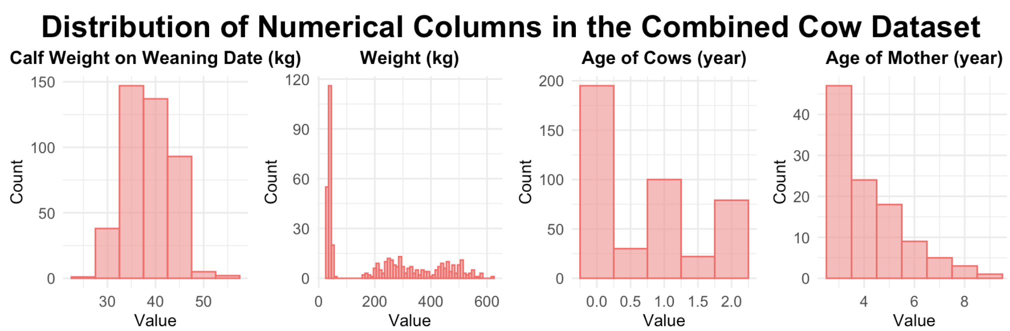

Figure 4.7: Histograms Showing the Distribution of Numerical Columns in the Combined Cow Dataset, Including Animal Factors (Calf Weight on Weaning Date, Weight, Age of Cows, Age of Mother)

Figure 4.7 presents the histogram distribution of cow data. In terms of the calf weight on the weaning date, this factor seems to have a normal distribution, with most calves weighing between 30 and 40 kg. Regarding weight, the distribution shows a wide range, indicating variability in cow sizes, but most weights are clustered at the beginning. This clustering might be because weights are primarily recorded on the weaning day, suggesting that farmers should record weights throughout the growth period for more comprehensive insights. Regarding the age of cows, this factor ranges between 0 and 2 years, indicating a young population in the combined cow dataset as the data recording is mostly done on weaning weight (highest histogram recorded at age 0.0 around 200 cows). In terms of mother's age, this factor shows a right-skewed distribution, with most mothers aged between 2 and 5 years.

Growth Trends of Cows

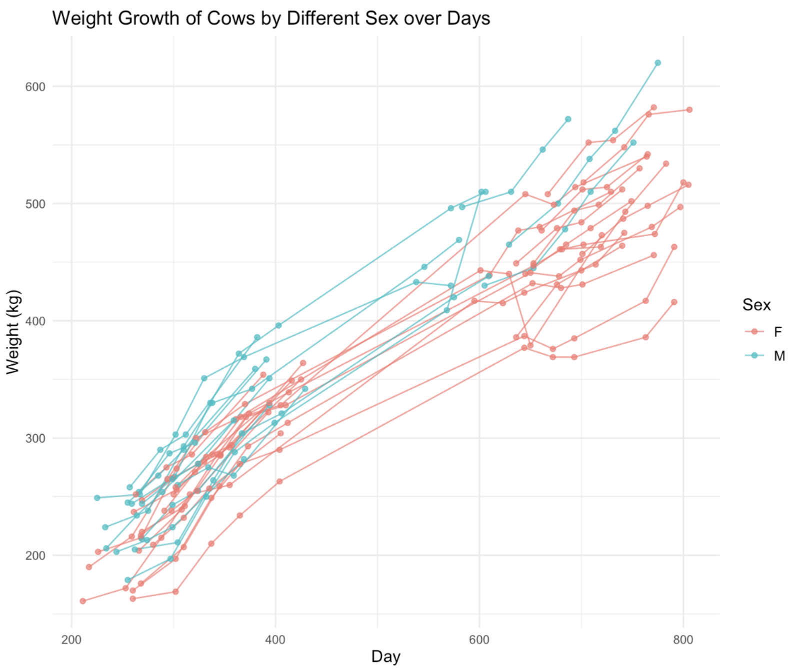

Figure 4.8: Weight Growth Trajectories of Cows Over Days for Cows Recorded Born in Different Years from 2014 to 2024, Differentiated by Sex (Male and Female)

The line graph in Figure 4.8 represents both male and female cows' growth trends, showing steady weight gain over time, with monitoring spanning around 800 days (more than two years). This highlights the importance of long-term monitoring in fully understanding growth dynamics. The insight from the graph is that male cows illustrate higher weight gain than female cows, with the difference becoming more clearly seen after approximately 400 days (around one year and one month).

However, Figure 4.8 provides some limitations to the recording of data. The broken lines suggest incomplete or inconsistent recording for some cows. Furthermore, the varying lengths of data recording, with some cows being tracked for around 800 days while others for less than 400 days, indicate a need for a standardised weight measurement process. To address the problem of recording the data, the farmer might establish consistent recording data practices to ensure the farm's progress in standardised weight measurement.

Correlation of the Combined Cow Datasets

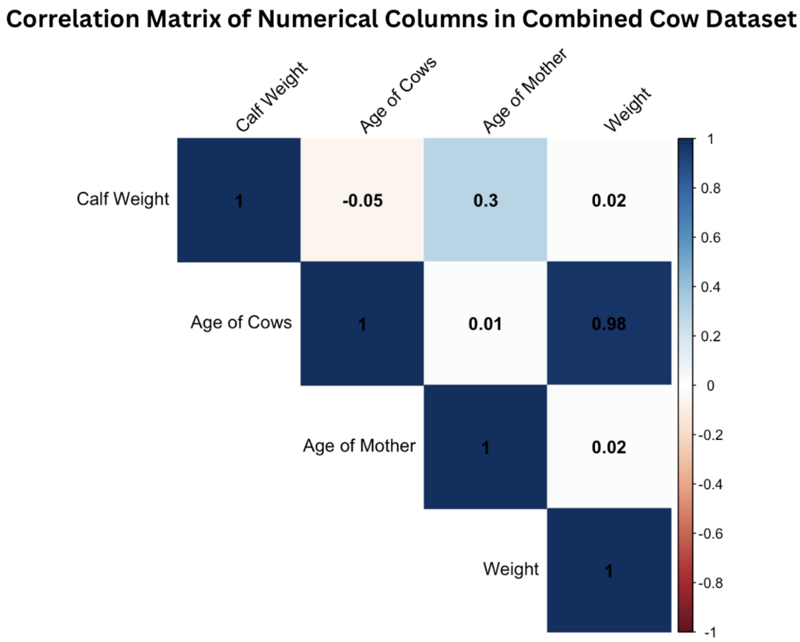

Figure 4.9: Correlation Matrix of Numerical Columns in the Combined Cow Dataset (Blue indicates a positive relationship, while red indicates a negative relationship)

The correlation matrix of numerical columns in the combined cow dataset, as in Figure 4.9, provided several key insights: There is a moderate positive correlation (0.3) between calf weight and the age of the mother, suggesting that older mothers tend to have heavier calves. However, this insight could be opposed by the insight from the scatter plot in Figure 4.3(b) that suggests no clear relationship between the mother's age and the weaning weight of calves.

Moreover, the age of cows shows a very strong positive correlation (0.98) with weight, indicating that older cows are commonly heavier than the younger.

However, the data contain some weak correlations, such as the very weak positive correlation (0.02) between calf weight and cow weight, suggesting that the calf's weight at the beginning might not affect the cow's growth. Furthermore, the age of the mother also has a very small correlation (0.02) to the weight of the cow, meaning that the age of the mother might not affect the growth of the cows. This insight could be rechecked through the analysis by fitting the LMMs to understand the coefficient of each factor in Chapter 5.

4.3 Summaries the Main Insight from the Statistical Summary and EDA

Statistical summary and EDA of the lamb and cow datasets, along with weather data, provide some key insights that could be used to report to Crag House Farm and help farmers improve their management practices.

In terms of lambs, the birth rate of female lambs each year is generally higher than that of male lambs, with pedigree lambs being more prevalent than commercial lambs, except in 2017. Furthermore, male lambs have statistically significantly higher weaning weights than females. Additionally, single-born lambs are the statistically significant heaviest weight compared with twins and triplets, as shown by the statistical test (independent t-test).

Regarding cows, The growth trends for cows show that both sexes gain weight steadily over time, with males generally having higher weights than females. Moreover, the mother's age might not impact the weight of calves from the non-clear relationships of these two factors from the scatter plot in Figure 4.3(b).

As regards weather conditions, these could impact the farm animals' growth from the variability of the weather conditions and extreme weather conditions such as high wind speed or high humidity, might cause stress to the farm animals, affecting the growth rate of the farm animals.

Based on these insights, recommendations for farmers could enhance animal welfare and management practices at Crag House Farm, potentially leading to higher productivity and profitability. For example, prioritising the feeding of animals with high weaning weights could lead to heavier adult animals, improving overall growth outcomes. Additionally, investing in protective structures to shield animals from extreme weather conditions may help mitigate the negative impacts of environmental factors on lamb health and growth.

Chapter 5

Longitudinal Data Analysis and Result

This chapter presents the fitted LMM to analyse the longitudinal data of farm animals to achieve objective 1, which is finding the factors that affect the farm animals' growth, and objective 2, which is to tune the model to improve its accuracy and robustness. An explanation of each analysis will be provided in the paragraphs below. The analysis will be done on two main animals of Crag House Farm which are lambs and cows.

Explanation of First Analysis to Achieve Objective 1

The first analysis will be done to achieve objective 1 of the study to find the factors that affect farm animal growth. The analysis involves fitting the data using a linear mixed model, as in Equation 2.1, by setting the fixed effects and the possible factors across all observations that influence the growth of animals, such as sex, age, and weather conditions. These factors are already detailed in Sections 3.3 of the data preparation process and illustrated the all factor that used in this analysis in Figures 3.1 and 3.2. The fixed effects after fitting the model through R programming will provide the resulting coefficients (β) that could be used to interpret the magnitude and statistical significance of various factors affecting farm animal growth and help in understanding how each variable impacts the growth rates of the animals.

In terms of the random effect of LMM, u_i is used to handle the random effect of variability among individual farm animals from the repeated measurements, which are longitudinal structures as in Figures 3.1 and 3.2 identified by repeated measurement of the same lamb ID and calves ID on different ages. The u_i term accounts for the variability of individual animal' weight that is not covered by the fixed effects. This term might capture the unobserved characteristics of each animal, allowing for a more accurate and nuanced understanding of the farm animals' growth patterns.

Explanation of Second Analysis to Achieve Objective 2

The second analysis will be conducted to achieve objective 2, which involves tuning the model to improve its accuracy and robustness. Additionally, the tuned model will be compared with the original model from objective 1 to interpret the results and identify any changes in the parameters after tuning. This analysis will use the feature selection technique to tune the model because, from the model analysis conducted in Section 5.1.1 and 5.2.1, it is evident that the model includes many input variables affecting weight. While some variables are statistically significant, others are not. Having too many input variables can lead to the problem of overfitting, where the model performs well on the training data but poorly on new or unseen datasets (Subramanian and Simon, 2013; Ying, 2019). Reducing the number of parameters by focusing on the most important ones is necessary to improve the model's generalisability and performance on different datasets (Ying, 2019). Therefore, this section uses the feature selection technique to reduce some non-effect parameters to reduce the number of parameters.

Moreover, this analysis involves using cross-validation, where the data is randomly split into training and validation sets multiple times, with both the LMM models for lambs and cows used in the cross-validation process (Maleki *et al.*, 2022). This method is integrated with the feature selection technique by randomly selecting the input sets to fit the model to determine which models with different sets of input factors could best suit the data by recording AIC value through the times of cross-validation to calculate the average AIC value of models, which is a measure of how good the model is prediction by penalising the number of parameters k as in Equation 2.3 to prevent overfitting (Leiter and Turner, 2017; Toga, 2015).

The models with the top 5 lowest AIC values are selected for further analysis. The top five models with the lowest AIC values are compared based on the RMSE criteria (comparing the difference of actual values to the prediction value from LMMs) to find the model with the lowest average RMSE, which indicates the highest predictive accuracy (Pandey *et al.*, 2022). This process ensures that the final models not only fit the data well and have high accuracy but also generalise effectively to new data, providing reliable predictions for the weight of farm animals. The final model will be interpreted by examining the coefficients to ensure that the results from the analysis in Section 5.1.2 and 5.2.2, which used all input variables, and the model after tuning are aligned and consistent.

5.1 Longitudinal Analysis of Lambs

5.1.1 Analysis to Achieve Objective 1, Focusing on Lambs: Identifying Factors Affecting Farm Animal Growth from All Possible Factors in the Dataset

Factors Affecting Lamb Growth with Result from R Programming

The code to fit the LMM model in R is:

```
    model <- lmer(Weight ~ Age + Sex + Type + WeaningWeight
+ SireID + AgeOfMother + BirthType + WeightSeason + WindSpeed
+ Temperature + GlobalRadiation + Humidity + (1 | ID), data = lamb)
```

Where The term `(1 | ID)` in the model represents the random effect that accounts for variability due to repeated recordings of the same lambs' IDs.

After fitting the model in R, the coefficient in the β that stands for the fixed effect in the

Table 5.1. The coefficient that is received from the programming could be used to interpret the statistical significance by calculating the 95% confidence interval of the coefficient through Equation 5.1.

$$\mathrm{CI}_{95\%} = \text{coefficient} \pm \left(t_{\alpha/2,\mathrm{dof}} \times \text{Std. Error}\right) \tag{5.1}$$

where

- $\mathrm{CI}_{95\%}$ is the 95% confidence interval.
- $t_{\alpha/2,\mathrm{dof}}$ is the critical value from the t-distribution for a given significance level α and degrees of freedom.
- Std. Error is the standard error of the coefficient.

Factor	**Coefficient (kg)**	**Standard Deviation (kg)**	**Lower bound of 95% Confidence Interval (kg)**	**Upper bound of 95% Confidence Interval (kg)**
(Intercept)	-11.25	3.35	-17.63	-4.88
Age	24.47	0.81	22.87	26.02
Sex (Male)	0.36	0.54	-0.65	1.36
Type (Pedigree lambs)	-1.41	0.88	-3.06	0.25
Weaning weight	0.82	0.05	0.72	0.92
Sire (ID 590)	-2.11	2.16	-6.15	2.00
Age of mother	1.21	0.71	-0.13	2.55
Twin birth lambs	-0.07	0.72	-1.44	1.29
Triplet birth lambs	-0.72	1.08	-2.74	1.30
Season weight recorded (spring)	-0.81	0.67	-2.14	0.49
Season weight recorded (summer)	-6.23	0.48	-7.16	-5.31
Season weight recorded (winter)	-2.53	0.51	-3.52	-1.54
Wind speed	0.00	0.00	-0.01	0.00
Temperature	-0.17	0.05	-0.27	-0.08
Global radiation	0.03	0.00	0.02	0.04
Humidity	0.17	0.02	0.12	0.22

Table 5.1: Coefficients, Standard Deviation, and 95% Confidence Intervals for Fixed Effects that Influenced the Weight of the Lambs through LMM for Longitudinal Analysis

The fitted LMM model of longitudinal lambs analysis is:

$$
\begin{aligned}
\text{Weight}_{\text{individual lamb}} = & -11.25 + (24.47 \times \text{Age of lambs}) + (0.36 \times \text{Sex}_{\text{Male}}) - (1.41 \times \text{Type}_{\text{Pedigree}}) \\
& + (0.82 \times \text{Weaning Weight}) - (2.11 \times \text{Sire}_{IDno.590}) + (1.21 \times \text{Mother Age}) \\
& - (0.07 \times \text{BirthType}_{\text{Twin}}) - (0.72 \times \text{Birth Type}_{\text{Triplet}}) \\
& - (0.81 \times \text{Season Recorded}_{\text{Spring}} - (6.23 \times \text{Season Recorded}_{\text{Summer}}) \\
& - (2.53 \times \text{Season Recorded}_{\text{Winter}}) - (0.17 \times \text{Temperature}) \\
& + (0.03 \times \text{Global Radiation}) + (0.17 \times \text{Humidity}) + u_{\text{ID}} + \epsilon
\end{aligned} \tag{5.2}
$$

The LMM of lamb growths Equation 5.2 is the equation that uses the LMM general equation as Equation 2.1 by including the terms of fixed effects from Table 5.1. Moreover, the random effect variable, which is $u_{\text{ID}} \sim \mathcal{N}(0, \sigma^2_{\text{ID}})$, is the random effect for repeated measurement of the same lamb ID in this model are included of 70 lambs, and is normally distributed with mean zero and variance σ^2_{ID}. Furthermore, the term residual error is also included ($\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon)$), which is normally distributed with mean equal to 0 and variance σ^2_ϵ.

The random effect and the residual error in this longitudinal model as Equation 5.2 are defined as $u_{\text{ID}} \sim \mathcal{N}(0, 2.64)$ and $\epsilon \sim \mathcal{N}(0, 5.32)$, where $\sigma^2_{\text{ID}} = 2.64$ and $\sigma^2_\epsilon = 5.32$ are the variances for the random effect and the residual error that received from the programming, respectively.

Interpretation of the Results (Lamb Data) and Summary of Insights for Farmers

To interpret the result from the Table 5.1 and Equation 5.2 in terms of the fixed effects (β), several statistically significant factors influencing lamb weight (measured in kilograms) are found. However, there are still some factors that do not provide a statistically significant effect on the growth of lambs. These findings are summarized in the following sections, beginning with the intercept term, followed by statistically significant positive impacts, statistically significant negative impacts, and factors that do not provide statistically significant effects.

The intercept term represents the baseline weight when all other factors are at their reference levels. For example, Age is set to zero, and the reference category for Sex is female. The estimated baseline weight is -11.25 kg, with a 95% confidence interval ranging from -17.63 to -4.88 kg. This negative intercept suggests that, without any significant positive factors contributing, the baseline weight is substantially lower.

Positive statistically significant factors influencing the increase in lamb weight are age, weaning weight and humidity. In terms of age, the age of the lamb has a highly significant positive effect on weight, with a coefficient of 24.47 kg per year of age. The 95% confidence interval (from 22.87 to 26.02 kg) supports the statistical significance of this factor. So, farmers might keep the lambs for a longer duration to gain more weight. Regarding weaning weight, this factor has a statistically significant impact on future growth. Each additional kilogram at weaning could increase the future weight by 0.82 kg, supported by a positive range of confidence between 0.72 and 0.92 kg. This insight could inform farmers that focusing on early nutrition for high weaning weight is crucial, as higher weaning weights could translate to better growth. As regards the weather conditions effect, global radiation has a small but significant positive effect on lamb weight, with an increase of 1 watts per square meter of global radiation, which could improve the weight of lamb by around 0.03 kg. This effect can be explained by the role of global radiation in enhancing plant growth. Higher levels of global radiation promote a more efficient plant photosynthesis system, which results in healthier and growing crops (Trnka *et al.*, 2007). These improved crops provide better-quality feed for the lambs, leading to improved nutrition of the farm animals and, consequently, better weight gain of lambs. Additionally, for the weather conditions effect, humidity also has a small but significant positive effect on lamb weight, with an estimated impact of 0.17 kg per percentage increase in humidity. The advantage of humidity control at high levels could be supported by the fact that higher humidity levels promote better grass growth by maintaining adequate moisture in the soil, which sustains high-quality food for lambs (Ford and Thorne, 1974). Therefore, for suggestion farmers could provide shelter that can control humidity levels to maintain a high level, which positively affects lambs' growth and sustains high-quality food for lambs.

Negative statistically significant factors influencing the decrease in lamb weight are the seasonal effect and temperature. Regarding the seasonal effect, lambs weighed in summer and winter have statistically significantly lower weights compared to autumn support by both season confidence intervals do not include the 0. Summer lamb and winter lamb weigh around 6.23 kg and 2.53 kg, respectively, lower than autumn lamb. This insight reflects the challenges of maintaining growth during extreme seasonal conditions. To advise the farmer, the farm might need seasonal management strategies to mitigate these negative effects, such as improving feed and shelter during extreme seasonal conditions. As regards temperature, rising temperatures have a small but statistically significant negative impact on lamb weight. Increasing it to 1 degree Celsius might reduce the animal's weight by around 0.17 kg. This could suggest to the farmer to provide cooling strategies during hotter months that could help maintain the growth rates of lambs.

The factors that are not statistically significant for lamb growth include the male sex, pedigree type, sire (the father of the lambs), the birth type of lambs (twin, triplet), the seasonal effects of spring, Global radiation and wind speed. Regarding the male lambs, the EDA analysis in Section 4.2.1 in Figure 4.5(a) from the growth trends of the lambs suggests the male lambs might have higher weight compared to the females, supported by the coefficient of LMM, which suggests males might have approximately 0.36 kg higher than females. However, the LMM result suggests that male lambs are not statistically significantly higher compared to the females from the confident interval range between -0.65 and 1.36 kg indicating that sex should not be a primary focus for managing lamb weight. In terms of pedigree lambs, the model suggests that they weigh less than commercial lambs; however, this difference is not statistically significant, with the confidence interval ranging from -3.06 to 0.25 kg. Nonetheless, this finding contrasts with the farmer's information that commercial lambs should weigh significantly more, as this type is bred for meat production, while pedigree lambs are primarily for maintaining breed lines and may not need to be as heavy. Therefore, this discrepancy will be re-examined in the analysis of the tuned model, as discussed in the second objective in Section 2.2.2. In terms of sire, the influence of sire ID 590 on weight is not statistically significant, as the confidence interval (-6.15 to 2.00 kg) includes zero, indicating the selection of breeder is also not the primary focus for managing lamb weight. Regarding the birth type, twin and triplet lambs might decrease their weight by around 0.07 and 0.72 kg, respectively, compared to the reference type, which is single, but this effect is not significant. This insight could be supported by the EDA that did independent t-tests to compare different weaning weights of different birth types of lambs in Figure 4.2 that suggests the singles is highest weaning weights compared to twins and triplets and might grow as the highest weight from the coefficient of LMM indicates. However, this insight is not statistically significant; the farmer should focus on general care rather than birth-type-specific interventions. In terms of wind speed, the LMM coefficient of 0 suggests not having any effect on the lambs.

To interpret the result of random effect and residual errors u_{ID} and ϵ terms in Equation 5.2, the random effect, u_{ID}, accounts for the variability between different groups (different Lamb IDs) that are not captured by the fixed effects. This term allows the model to handle data that is grouped (repeated record of individual lambs), ensuring that each individual lamb can have its own baseline weight. The random effect is modelled as normally distributed with a mean of zero and a variance of 2.64. This could suggest that while the average effect across all individual lambs is zero, individual lambs may deviate from this average of around 1.63. In terms of the residual error term, ϵ, this term captures the remaining variability in weight that might not be explained by either the fixed effects or the random effects. This term is also modelled as normally distributed with a mean of zero and a variance of 5.32, which suggests individual lambs may deviate from this average of around 2.31 kg.

To conclude, by understanding these factors, farmers could make informed decisions to improve lamb growth and overall farm profitability. This includes prioritising management practices that improve the positive factors and mitigate negative factors, which could lead to more efficient and profitable operations at Crag House Farm.

5.1.2 Analysis to Achieve Objective 2, Focusing on Lambs: Tuning the Model to Achieve a High-Accuracy and Robust Model

Tuning the LMM of the Lambs

The cross-validation process compares the model with the feature selection technique in lambs data. The fixed effect input feature could be based on 12 input features: age, sex, type, weaning weight, sire, age of mothers, birth types, the season of weight recording, average wind speed, temperature, global radiation, and humidity on the weight recording day.

The tuning used the cross-validation process to compare the model with the feature selection technique by randomly selecting to create the list of various models with different inputs, which is shown in the example in Figure 5.1. Figure 5.1 illustrates that the randomly selected process in R provides 4095 different models of LMM for lambs with different input lists.

Name	Type	Value
formulas	list [4095]	List of length 4095
[[1]]	formula	Weight ~ Type + (1 | ID)
[[2]]	formula	Weight ~ WeaningWeight + (1 | ID)
[[3]]	formula	Weight ~ Age + (1 | ID)
[[4]]	formula	Weight ~ WeightSeason + (1 | ID)
[[5]]	formula	Weight ~ Temperature + (1 | ID)
[[6]]	formula	Weight ~ Humidity + (1 | ID)
[[7]]	formula	Weight ~ Sex + (1 | ID)
[[8]]	formula	Weight ~ SireID + (1 | ID)
[[9]]	formula	Weight ~ AgeOfMother + (1 | ID)
[[10]]	formula	Weight ~ BirthType + (1 | ID)

Figure 5.1: The Example List of Different Input Formulas Captured From R to Compare in the Cross-Validation Process of AIC Value of Different LMM of Lambs.

This was followed by Running the 20-time cross-validation through the 4095 LMMs of lambs and recording the average AIC in 20 times of CV to compare the model's performance. Additionally, the top 5 models with the lowest AIC, which suggests the model has a good balance of fitting the data with the penalise of the adding of the input parameter, as illustrated in Table 5.2 in the second column.

Following the process, conduct cross-validation 100 times to compare the accuracy of the top five models with the lowest average AIC values by using RMSE to evaluate. The RMSE for each model fitting is recorded during each iteration. The model with the lowest average RMSE is identified as the best model for predicting the weight of lambs with the highest accuracy as in the third column of Table 5.2.

Table 5.2 shows the model with the lowest average RMSE value in this analysis contains two models, **Model 1** and **Model 2**, in Table 5.2, that have the same average RMSE of 2.492 kg. Moreover, the average of the AIC value of the two models is also equal. Therefore, the reason to choose the best model between the two models which includes the type of lambs or not is used from the insight of the model that was analysed in Section 5.1.1 and the assumption that received from the farmer information. To elaborate, the model before tuned indicates that pedigree lambs are not statistically significantly different in weight compared to commercial lambs, which contrasts with the farmer's assumption that commercial lambs should weigh significantly more. To resolve this conflict, it is crucial to interpret the tuned model that includes the type factor. Therefore, **Model 2** from Table 5.2, which includes the type variable, may be better at capturing this relationship and could be preferred for predicting lamb weight.

The best LMM after tuning is **Model 2** in Table 5.2 is:

```
Weight ~ Age + Type + WeaningWeight + SireID + AgeOfMother + WeightSeason
+ Temperature + GlobalRadiation + Humidity + (1| ID)
```

The model selection input	**1st CV to record average AIC values**	**2nd CV to recorded average RMSE (kg)**
`Model 1: Weight ~ Age + WeaningWeight + SireID + AgeOfMother + WeightSeason + Temperature + GlobalRadiation + Humidity + (1|ID)`	1672.203	2.492
`Model 2: Weight ~ Age + Type + WeaningWeight + SireID + AgeOfMother + WeightSeason + Temperature + GlobalRadiation + Humidity + (1|ID)`	1672.203	2.492
`Model 3: Weight ~ Age + WeaningWeight + SireID + AgeOfMother + BirthType + WeightSeason + Temperature + GlobalRadiation + Humidity + (1|ID)`	1672.568	2.497
`Model 4: Weight ~ Age + Type + WeaningWeight + SireID + AgeOfMother + BirthType + WeightSeason + Temperature + GlobalRadiation + Humidity + (1|ID)`	1672.568	2.497
`Model 5: Weight ~ Age + Sex + Type + WeaningWeight + SireID + AgeOfMother + BirthType + WeightSeason + Temperature + GlobalRadiation + Humidity + (1|ID)`	1673.352	2.495

Table 5.2: The List of Top Five Models With the Lowest Average AIC Value During the 20-Time Cross-Validation of Lamb Data. Moreover, the Second Step of 100 Times Cross-Validation Was Done on Only the Top 5 Models to Record the RMSE. (The Term `(1|ID)` in the Model Represents the Random Effect That Accounts for Variability Due to Repeated Recordings of the Same Lambs' IDs.)

This LMM including the 9 fixed effect input features that include the type, weaning weight, age, the season that weight was recorded, sire (father of lambs), age of the mother, temperature, global radiation and humidity is the model with the lowest average RMSE value at 2.492 kg, suggesting this model could be the best model with the balance of fit the data well and have high accuracy in predicting the weight of lambs. Regarding ensuring accurate predictions of lamb weight, farmers at Crag House Farm should carefully record the 9 input factors by carefully recording the type of lamb and weight on the weaning day. Furthermore, farmers should regularly update the age of lambs and also consider the age of the breeder (mother). Additionally, the specifics of the sire (father) should also be carefully recorded. Moreover, when weighing the date, the farmer could record the date and season. Additionally, farmers should record environmental conditions like temperature, global radiation and humidity by using reliable instruments to receive a high accuracy of measure. Consistent recording practices could enhance the data reliability and predictive accuracy of the LMM models.

In terms of interpreting the average RMSE value of 2.492 kg, this value suggests that the model might be suitable for real-world scenarios to predict lamb weights. This RMSE value suggests the error of prediction of LMM at approximately 2.5 kg. This level of error might be considered acceptable for farming operations that do not require extremely precise weight measurements. Therefore, this value indicates that the model is reasonably accurate and could be used to predict lamb weights effectively, providing a useful tool for real-world application of farm management decisions. However, it's important for a farmer on the Crag House Farm to consider their specific needs and the acceptable margin of error for their farm operations.

Comparison of the Accuracy of the LMM Before and After Tuning

To compare the accuracy of the tuning LMM and the model before tuning, this process is done on the paired t-test which records the RMSE of the tune and non-tune model at the 100 times of cross-validation because the paired sample t-test is commonly employed to compare the difference between two population means in a design where samples are matched or related (Rietveld and Van Hout, 2017). To elaborate, the paired t-test in this study used to compare based on the same lamb data is measured under two different conditions, which are before and after tuning LMM. By using the hypothesis that the average RMSE before and after tuning the LMM is similar, the paired t-test will evaluate whether there is a statistically significant difference between the average RMSE of the two conditions. By setting the null hypothesis as

- **Null Hypothesis ($\mathbf{H_0}$)**: The average RMSE before tuning is equal to the average RMSE after tuning.

$$H_0 : \mu_{\text{before}} = \mu_{\text{after}}$$

- **Alternative Hypothesis ($\mathbf{H_1}$)**: The average RMSE before tuning is not equal to the average RMSE after tuning.

$$H_1 : \mu_{\text{before}} \neq \mu_{\text{after}}$$

The result after the paired t-test of the model rejects the null hypothesis H_0 that the average RMSE after tuning is equal to the RMSE of the LMM before tuning. This finding demonstrates that the model after tuning provides a statistically significant difference between the RMSE before and after tuning, as indicated by an improvement in accuracy. The 95% confidence interval for the difference in RMSE is between 0.005 and 0.013 kg (degrees of freedom = 99, $p < 0.001$). Therefore, the evidence supports the assumption that the model after tuning, using the feature selection technique, is better than the old model.

Interpret the Changing of Tuning LMM of Lambs

To compare the coefficient of the LMMs model of Table 5.1 and Table 5.3 to find the changing insight after tuning the model it could be found that:

After tuning the LMM for lamb weight, 2 main key improvements were observed in terms of the type of lambs and the age of the mother, making this model more reliable compared to the previous one, as in Section while the other variable stay at the same way with changing a bit of magnitude. One of the most notable changes is that the type of lambs, which is pedigree, became statistically significant after tuning, indicating that pedigree lambs have a significantly lower weight compared to commercial lambs around 0.3 to 3.2 kg. This insight, after tuning, could be used to confirm the information the farmer told that the commercial lambs should grow better than the pedigree lambs because commercial lambs are bred for higher growth rates to produce meat production, while pedigree lambs are often maintained for breeding purposes. Therefore, focusing on breeding commercial lambs for higher growth rates could enhance farm productivity.

Additionally, the factor of the mother's age also turned statistically significant, demonstrating a positive effect on lamb weight. This suggests that older mothers may have a positive influence on the growth rates of their offspring. By considering the age of the mother as a factor in breeding decisions, farmers in Crag House Farm may improve the overall productivity of their farms, leading to better growth outcomes for lambs.

Moreover, the non-significant parameters have been cut from the model, which is Sex (Male), birth type and average daily wind speed. These three insights do not provide statistical significance in the model after the tune, as a model in Section 5.1.1.

Meanwhile, the other variables retained their effects, with only slight changes in their magnitude. This insight could ensure that the overall structure of the model remained consistent while providing more refined predictions.

Factor	**Coefficient (kg)**	**Standard Deviation (kg)**	**Lower bound of 95% Confidence Interval (kg)**	**Upper bound of 95% Confidence Interval (kg)**
(Intercept)	-13.62	2.70	-18.82	-8.43
Age	24.76	0.76	23.27	26.21
Type (Pedigree lambs)	-1.75	0.75	-3.18	-0.32
Weaning weight	0.85	0.05	0.75	0.93
Sire (ID 590)	-2.54	2.11	-6.60	1.50
Age of mother	1.50	0.54	0.40	2.50
Season weight recorded (spring)	-0.58	0.64	-1.85	0.66
Season weight recorded (summer)	-6.09	0.46	-7.00	-5.20
Season weight recorded (winter)	-2.40	0.50	-3.36	-1.43
Temperature	-0.18	0.05	-0.28	-0.09
Global radiation	0.03	0.00	0.02	0.03
Humidity	0.18	0.02	0.14	0.23

Table 5.3: Coefficients, the Standard Deviation, and 95% Confidence Intervals for Fixed Effects That Influenced the Weight of the Lambs Through Tuned LMM for Longitudinal Analysis of Lambs.

In summary, the tuned model could offer better statistical insights that align more closely with real-world scenarios. The improvement in the statistical significance of important variables, such as the type of pedigree lambs and the age of the mother, in which the improvement of the factor of type is linked with the information provided by farmers, could be used to support the assumption that the tuned model provides better insight than the old model. Therefore the tuned model provides a clearer understanding of the factors affecting lamb weight, making it a more effective tool for decision-making and management strategies to the farmer.

5.2 Longitudinal Analysis of Cows

5.2.1 Analysis to Achieve Objective 1, Focusing on Cows: Identifying Factors Affecting Farm Animal Growth from All Possible Factors in the Dataset

Factors Affecting Lamb Growth with Result from R Programming

After fitting the LMM, the coefficient from Equation 2.1 in terms of β that stands for the fixed effect and the u_i terms that handle the random effect obtained from programming are represented in Table 5.4 could be used to interpret the result and find the factor that statistically significant effect the growth of cow.

Factor	**Coefficient (kg)**	**Standard Deviation (kg)**	**Lower bound of 95% Confidence Interval (kg)**	**Upper bound of 95% Confidence Interval (kg)**
(Intercept)	-34.39	25.53	-83.69	14.01
Age	236.14	3.19	230.09	242.48
Sex (Male)	5.98	5.75	-4.93	17.00
Weaning Weight	2.24	0.71	0.90	3.60
Season weight recorded (spring)	-4.52	6.33	-16.98	7.68
Season weight recorded (winter)	-7.60	6.32	-20.01	4.60
Sire (name Superman)	-15.89	19.72	-53.30	21.24
Sire (name Topgun)	13.93	5.87	2.79	25.24
Sire (name Unforgetab-ull)	-2.92	10.73	-23.42	17.72
Age of the mother	-1.64	1.92	-5.32	2.02

Table 5.4: Coefficients, the Standard Deviation, and 95% Confidence Intervals for Fixed Effects That Influenced the Weight of the Cows Through LMM for Longitudinal Analysis of Cows.

The fitted LMM model of longitudinal cows analysis is:

$$
\begin{aligned}
\text{Weight}_{\text{individual cow}} = & -34.39 + (236.14 \times \text{Age of cows}) + (5.98 \times \text{Sex}_{\text{Male}}) + (2.24 \times \text{Weaning Weight}) \\
& - (4.52 \times \text{Season Recorded}_{\text{Spring}}) - (7.60 \times \text{Season Recorded}_{\text{Winter}}) \\
& - (15.89 \times \text{Sire(name)}_{Superman}) + (13.93 \times \text{Sire(name)}_{Topgun}) \\
& - (2.92 \times \text{Sire(name)}_{Unforgetabull}) - (1.64 \times \text{Mother Age}) + u_{\text{ID}} + \epsilon
\end{aligned}
\tag{5.3}
$$

The LMM of cow growths as Equation 5.3 is the equation that uses the LMM general equation as Equation 2.1 by including the terms of fixed effect from Table 5.4. Additionally, the random effect variable of the cow data is also included in the LMM model, which is cow ID, which is the random effect for repeated measurement of the same cow ID. This random effect variable is normally distributed as $u_{\text{ID}} \sim \mathcal{N}(0, \sigma^2_{\text{ID}})$ with mean zero, and variance at σ^2_{ID}. Furthermore, the term residual error also includes $\epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon})$, which is normally distributed with a mean equal to 0 and variance at σ^2_{ϵ}.

By, the random effect and the residual error in this longitudinal model as Equation 5.3 are defined as $\mu_{\text{ID}} \sim \mathcal{N}(0, 304.60)$ and $\epsilon \sim \mathcal{N}(0, 553.70)$, where $\sigma^2_{\text{ID}} = 304.60$ and $\sigma^2_{\epsilon} = 553.70$ are the variances for the random effect and the residual error that received from the programming of the longitudinal analysis of cow model, respectively.

Interpretation of the Results (Cow Data) and Summary of Insights for Farmers

To interpret the results from Table 5.4 and Equation 5.3 in terms of the fixed effects, some factors are identified as statistically significant in influencing cow weight which are only positive effects, while some of the factors do not suggest a statistically significant effect on the growth of cows.

In terms of the intercept term, which is -34.39 kg, this value represents the baseline weight when all other factors are at reference levels (for example, age is a numerical factor at 0 years old and sex is the reference category, which in this study is set as female).

Regarding the factor that positively and statistically significantly affects cow growth, age statistically significantly increases weight by approximately 236 kg per year, with a 95% confidence interval around 230 to 242 kg, which does not include the 0. This suggests rejecting the null hypothesis that cows of different ages have similar weights, indicating a significant impact of age on cow weight. This insight could suggest farmer that maintaining older cows is beneficial for the farmer, as their continued growth leads to increased overall weight, which could result in higher profitability through selling them. Regarding weaning weight, it statistically significantly affects cow growth, as indicated by an increase of approximately 2.24 kg per 1 kg of weaning weight increase, with a 95% confidence interval ranging from 0.90 to 3.60 kg, which does not include 0. This suggests that weaning weight has a statistically significant positive effect on cow growth. For farmers, this insight highlights the importance of managing weaning practices effectively, as higher weaning weights could lead to improved growth outcomes for the cows. In terms of sires (father), the Sire name Topgun suggests a statistically significant impact by this sire could provide more than 13.93 kg than the reference sire, whose name is Nitro. This insight could be used to recommend the farmer to select the proper breeder to enhance calf weight.

As regards the factors which are not statistically significantly impact the weight of cows, the weight of male cows is around 6 kg more than females. However, this is not statistically significant from the 95% confidence interval, including 0. In terms of seasonal effects, these effects show that cows weigh less in spring and winter at approximately 5 and 8 kg, respectively compared to the reference season, which is autumn, but these effects are also not statistically significant. Among the sires, Topgun could be provided the calves that weigh significantly more than Superman and Unforgetabull which these two sires might not show any statistically significant effect. Regarding the mother's age, this factor has a slight negative impact on weight, but this effect is not significant from the confidence interval from -5.32 to 2.02.

To interpret the result of random effect and residual errors from equation 5.3, in terms of the random effect term (u_{ID}), this term accounts for variability among different cow IDs from the structure of longitudinal data, which is the structure of repeated measurement of the same cow ID to track the weight. This term is not captured by the fixed effects with a variance of $\sigma^2_{\text{ID}} = 304.6$ indicating that individual cows may deviate approximately 17.45 kg from the average weight. Regarding the residual error term (ϵ), this term could capture remaining variability not explained by the model, with a variance of $\sigma^2_{\epsilon} = 553.7$, suggesting individual cows might deviate around 23.53 kg from the average weight.

To conclude, understanding these factors could allow farmers to make informed decisions to improve cow growth. By focusing on significant factors such as age and specific sires, farmers can optimise the growth and health of their cows, leading to more efficient and profitable farming operations.

5.2.2 Analysis to Achieve Objective 2, Focusing on Cows: Tuning the Model to Achieve a High-Accuracy and Robust Model

Tuning the LMM of the Cows

The tuning process of the cow analysis is done on the same process as the lamb data by using the cross-validation process to compare the model with the feature selection technique in cow data. The fixed effect input feature could be based on six input features: age, the weight of calves on the weaning date, sex, the specific sire name or ID, the season the weight was recorded, and the age of the breeder(mother). These input factors were randomly selected as a set of predictors to fit the LMM model in the CV process. After randomly selecting to create the list of various models with different inputs, there are 63 different models of LMM for cows with different input lists, as in Figure 5.2.

The tuning process was followed by Running the 100-time cross-validation through the fitting of 63 LMMs of cows and recording the AIC in 100 times of CV to compare the average value of AIC that evaluate the model's performance. The results of CV to obtain the top 5 models with the lowest AIC are illustrated in Table 5.5 in the second column.

To elaborate, the reason the cow tuning process using cross-validation took more time with 100 iterations for 63 LMMs compared to the lamb data, which was used only 20 times, is due to the much larger number of models in the lamb dataset. The lamb dataset includes 4095 different models, requiring high performance of computer (CPU) to run. As a result, this study reduced the number of CV iterations to 20 for the lambs, while maintaining the standard 100 iterations for cows, which have around 7 times fewer models.

Name	Type	Value
formulas	list [63]	List of length 63
[[1]]	formula	Weight ~ Sex + (1 | CalfID)
[[2]]	formula	Weight ~ WeightAtSeason + (1 | CalfID)
[[3]]	formula	Weight ~ CalfWeight + (1 | CalfID)
[[4]]	formula	Weight ~ SireName + (1 | CalfID)
[[5]]	formula	Weight ~ AgeOfMother + (1 | CalfID)
[[6]]	formula	Weight ~ Age + (1 | CalfID)
[[7]]	formula	Weight ~ Sex + WeightAtSeason + (1 | CalfID)
[[8]]	formula	Weight ~ Sex + CalfWeight + (1 | CalfID)
[[9]]	formula	Weight ~ Sex + SireName + (1 | CalfID)
[[10]]	formula	Weight ~ Sex + AgeOfMother + (1 | CalfID)

Figure 5.2: The Example List of Different Input Formulas Captured From R to Compare in the Cross-Validation Process of the AIC Value of Different LMMs of Cows

Following the process, the next step is to conduct cross-validation 100 times to compare the accuracy of the top five models with the lowest average AIC values by using RMSE to evaluate the model accuracy. The model with the lowest average RMSE is identified as the best model for predicting the weight of cows with the highest accuracy which Table 5.5 in third column illustrates the result of the second times of CV in the third column average RMSE value (kg) recorded in the models.

The model which has the lowest average RMSE value from Table is:

`Weight ~ Age + WeaningWeight + WeightAtSeason + SireName + (1|CalfID)`

This best accuracy LMM includes the 4 fixed-effect input features that include age, the weight of calves on the weaning date, the season that the weight was recorded and the specific sire name is the model with the lowest average RMSE value at 24.826 kg, suggesting this model could be the best model with the balance of fitting the data well and being highly accurate in predicting the weight of cows.

To ensure accurate predictions of cow weight, farmers at Crag House Farm should carefully record the four key input factors by regularly updating the age of cows and recording the weight of calves at the weaning date. Additionally, documenting the weighting day is crucial to specifying the seasonal effect. Furthermore, farmers should record every sire ID or name to track the breeder effect of each lamb. Consistent recording practices could enhance the data reliability and predictive accuracy of the LMM models.

In terms of interpreting the average RMSE value of approximately 25 kg. This means the model's predictions deviate from the actual cow weights by around 25 kg, which might not be a large error because this error of predictive of animal weight might be larger in animals with higher weight. For example, Hereford cattle, a British breed primarily raised for meat production, have female cows typically reaching weights of up to 550 kg, while males can grow as heavy as 850 kg (Crawshaw, 2023). Therefore, an error of around 25 kg in predicted values might be considered acceptable when compared to the overall weight of the cows. However, determining whether this level of error is acceptable depends on the farmer's experience and decision-making needs. Farmers must consider if this accuracy meets their operational requirements to make an informed decision.

The model selection input	**1st CV to record average AIC values**	**2nd CV to recorded average RMSE (kg)**
`Model 1: Weight ~ Age + WeaningWeight + WeightAtSeason + SireName + (1|CalfID)`	2244.704	24.826
`Model 2: Weight ~ Age + Sex + WeaningWeight + WeightAtSeason + SireName + (1|CalfID)`	2239.740	24.908
`Model 3: Weight ~ Age + Sex + WeightAtSeason + SireName + AgeOfMother + (1|CalfID)`	2245.748	24.861
`Model 4: Age + WeaningWeight + WeightAtSeason + SireName + AgeOfMother + (1|CalfID)`	2242.343	24.884
`Model 5: Age + Sex + WeaningWeight + WeightAtSeason + SireName + AgeOfMother + (1|CalfID)`	2237.646	24.963

Table 5.5: The List of Top Five Models With the Lowest Average AIC Value During the 100-Time Cross-Validation of Lamb Data. Moreover, the Second Step of 100 Times Cross-Validation Was Done on Only the Top 5 Models to Record the RMSE. (The Term `(1|CalfID)` in the Model Represents the Random Effect That Accounts for Variability Due to Repeated Recordings of the Same Cows.)

Comparison of the Accuracy of the LMM Before and After Tuning

To compare the accuracy of the tuning LMMs of the cow and the model before tuning, this process is done on the paired t-test, the same as the lamb analysis, which uses the data to compare from the records the RMSE of the tune and non-tune model 100 times. By using the hypothesis that the average RMSE before and after tuning the LMM is similar, the paired t-test will evaluate whether there is a statistically significant difference between the average RMSE of the two conditions. By setting the null hypothesis the average RMSE of LMM after tune is equal to the average RMSE before tuning the model.

The result after the paired t-test of the model rejects the null hypothesis H_0 that the average RMSE after tuning is equal to the RMSE of LMM before tuning. This insight illustrates that the model after tuning provides a statistically significant difference between the RMSE before and after tuning by the accuracy of RMSE might improve from the support of t-test result with 95% confidence interval between in the range of 0.09 to 0.18 kg (degree of freedom = 99, $p < 0.001$). Therefore, the evidence supports the assumption that the model, after tuning by the feature selection technique, is better than the old model same as the lambs.

Interpret the changing of Tuning LMM of Lambs

Factor	**Coefficient (kg)**	**Standard Deviation (kg)**	**Lower bound of 95% Confidence Interval (kg)**	**Upper bound of 95% Confidence Interval (kg)**
(Intercept)	-34.39	25.53	-83.69	14.01
Age	236.14	3.19	230.09	242.48
Sex (Male)	5.98	5.75	-4.93	17.00
Weaning Weight	2.24	0.71	0.90	3.60
Season weight recorded (spring)	-4.52	6.33	-16.98	7.68
Season weight recorded (winter)	-7.60	6.32	-20.01	4.60
Sire (name Superman)	-15.89	19.72	-53.30	21.24
Sire (name Topgun)	13.93	5.87	2.79	25.24
Sire (name Unforgetab-ull)	-2.92	10.73	-23.42	17.72
Age of the mother	-1.64	1.92	-5.32	2.02

Table 5.6: Coefficients, the Standard Deviation, and 95% Confidence Intervals for Fixed Effects That Influenced the Weight of the Cows Through LMM After Tuning for Longitudinal Analysis of Cows.

There are some changes observed in Table 5.6 after tuning the LMMs which are:

In terms of seasonal effects, the winter season has changed to provide a statistically significant negative impact on weight, with the effect increasing from -7.60 kg to -10.08 kg, demonstrating the model's improved ability to account for seasonal variations. While other seasons still did not have a statistically significant effect on cow weight. This insight could provide farmers with a clearer view of when season cows are likely to gain or lose weight throughout the year.

Regarding the sire effect, the positive effect of the sire name Topgun, whose impact increased from 13.93 kg to 17.57 kg, highlights the updated model's improved sensitivity to genetic differences in sire selection. This adjustment provides more reliable guidance to farmers about which sires are likely to produce heavier offspring, thereby supporting better-informed breeding decisions.

Meanwhile, the other variables retained their effects, with only slight changes in their magnitude. This insight could ensure that the overall structure of the model remained consistent while providing more reliable predictions.

In conclusion, the tuned model could be considered more reliable because it offers refined estimates that better reflect the key influences on cow weight. By reducing the influence of less significant factors, such as the sex and age of the mother, the model offers more accurate predictions. This improved accuracy provides farmers with better data for decision-making, particularly concerning breeding strategies and seasonal management.

Chapter 6

Discussion and Conclusion

6.1 Discussion

6.1.1 The Discussion of Key Findings from the Study

Regarding the key findings from objective 1, the results from this study highlight several critical factors influencing the growth and weight of farm animals, particularly lambs and cows, at Crag House Farm in Leeds. The LMMs utilised in this study have provided insights into how various factors, such as age, weaning weight, birth type, sire (father of animal), and seasonal effects, contribute to weight variations. To elaborate, the age of the animals significantly impacts weight gain, with older lambs and cows showing a significant increase in weight compared to younger ones, consistent with the study of Gaskins *et al.* (2005), which suggests that older animals tend to weigh more than younger ones because they have had more time to accumulate body mass. Furthermore, this study also finds that higher weaning weights in lambs and cows may lead to significantly faster growth than lambs and cows with lower weaning weights. This finding could be supported by Hatcher *et al.*'s article (2008), which mentions the idea that weaning weight is crucial for determining post-weaning weight and growth rates, with its impact lasting up to 6 months. Although lighter weaners can experience compensatory growth with adequate nutrition, they are over twice as likely to die compared to heavier weaners (Hatcher *et al.*, 2008). This suggests that a higher weaning weight might be healthier than a low one. Moreover, the study of Simeonov *et al.* (2014) also mentioned that single lambs with higher birth weights demonstrated faster growth compared to the lower ones; however, no significant differences in growth were observed for twin lambs ($p > 0.05$).

Additionally, the birth type of animals, particularly in lambs for this study, was found to have a significant impact on their weaning weight from the EDA analysis in Section 4.1.1. For example, single-born lambs provided statistically significantly higher weaning weights than twins, triplets, and quadruplets. This finding is supported by the study of Assan and Makuza (2005), which highlights that lambs born as singles were significantly heavier at weaning compared to those born as twins across different sheep breeds, including Mutton Merino, Dorper, and Indigenous Sabi sheep. For example, in Indigenous Sabi sheep, single-born lambs were 0.78 kg heavier at weaning than twin lambs (Assan and Makuza, 2005). However, the impact on weight development from the birth type of animal throughout growth was not statistically significant in this study, suggesting that while birth type influences initial weight, it may have less effect on long-term growth trajectories of lambs, as supported by the article of Smith *et al.* (1982), which reported that single and twin calves have similar weights after growth at 1 year old. Additionally, this statement could be supported by the findings of De Rose and Wilton's (1991) study, which demonstrated that post-weaning growth was not statistically different for single and twin calves, with a p-value at around 0.5.

Moreover, the choice of the breeder (sire) was found to have a positive effect on animal weight in this analysis, especially in terms of calves, which was found statistically significant in this study, suggesting that the breeder named Topgun could have provided a higher cow growth rate compared to other sires. This statement could be linked to the study of Coleman *et al.* (2021), which highlights that different sires produced calves with varying birth weights, ranging from 33.3 to 41.4 kg, and pre-weaning growth rates, with average daily gains (ADG) ranging from 0.63 to 0.76 kg/day. Additionally, the study of Elbert *et al.* (2020) suggests that pigs from different sire lines show notable statistical differences in growth rate. For example, pigs from sire line A (Synthetic line) grew faster ($p < 0.01$) and reached a higher final body weight ($p < 0.01$) compared to those from line B (Piétrain line). This variation in weight from different breeders indicates that genetic factors associated with the sire play a crucial role in the growth performance of the animal offspring.

Additionally, this analysis also suggested considering the effect of seasonal conditions, which is particularly statistically significant for lambs in this study; the study found that lambs tend to lose weight in summer and winter when the weather might be extreme in that season, such as extreme wind or temperature fluctuations. By providing adequate shelter for the animals could protect animals from extreme weather conditions. Building shelters or improving existing ones could reduce stress on animals and create a more stable environment for growth, as supported by the study of Fisher (2007), which demonstrated that providing shelter to farm animals during extreme weather conditions significantly reduces thermal stress that occurs when animals cannot maintain a stable body temperature due to extreme weather conditions and that providing shelter could mitigate these effects. To elaborate, animals can choose to modify their behaviour to suit them, such as using the shelter to seek shade during hot conditions, thus reducing thermal stress. This reduction of thermal stress could improve growth rates and the overall health of animals, leading to good conditions for the growth of the animals.

In terms of objective 2 of this study, tuning the model to achieve a more accurate and robust prediction of animal weight was crucial to improving the model's accuracy. By tuning the LMMs through cross-validation and feature selection techniques, the study was able to find and compare the predictive accuracy of the models and improve the prediction accuracy. This study shows that the RMSE after-tune could statistically reduce the prediction error between 0.005

and 0.013 kg for lambs and between 0.02 and 0.18 kg for cows. This could be supported by the article of Hasan *et al.* (2015), which reported that using a linear regression-included feature selection method, their study was able to select a smaller, more relevant set of features, which led to higher classification accuracy than traditional methods using a larger set of features. Moreover, Karabulut *et al.* (2012) also suggest that feature selection can significantly enhance the performance of various classifier models, including Naïve Bayes, Artificial Neural Networks (like Multilayer Perceptron), and decision tree classifiers (like J48). By pre-processing datasets with feature selection methods, their study observed up to a 15.55% improvement in classification accuracy.

In terms of this longitudinal analysis study of farm animals, the results from the best robust and accurate model, after tuning through cross-validation combined with the feature selection technique, can be a valuable tool for farmers, allowing them to make data-driven decisions about management practices and interventions to help their animals reach target weights (predict weight) more effectively. Additionally, this process not only improved the reliability of the weight predictions but also provided a better understanding of the key factors that most significantly influence growth by making the updated model more reliable by cutting some of the less important factors. To elaborate, the final best accuracy models demonstrated this by cutting unnecessary input features: for lambs, excluding features such as sex, age of mother and wind speed resulted in an RMSE of around 2.5 kg, while for cows, removing features like sex and age of mother produced an RMSE of approximately 25 kg. However, these differences in the optimal input features for different farm animals suggest that different species may require distinct sets of significant predictors to model growth outcomes accurately.

6.1.2 The Discussion of Limitations of this Study

Regarding the limitations of this study, several key challenges were found through the analysis that might have impacted the results.

Firstly, this study is impacted by the limited availability and completeness of datasets, particularly due to the lack of continuous records of farm animals. The lamb dataset has a missing value of around 27% of total lamb data, and the cow dataset has a missing value of around 30%. This gap in data collection may have introduced biases and affected the reliability of the findings. To address these issues, future research might focus on enhancing data quality by ensuring more consistent and comprehensive data collection. This could be done through the use of automated systems that can maintain continuous and accurate records over time. This approach might improve the accuracy of the analysis by reducing the biased effect of the missing value.

Additionally, the absence of weather condition data, such as wind speed, temperature, global radiation, and humidity, in the cow growth dataset meant that the impact of these environmental factors could not be evaluated for cows due to the weather condition data covered only until 2020 while the weighting of cow records started after 2020. This limitation could be addressed by sourcing reliable weather condition data, such as wind speed, temperature, global radiation, and humidity, from trusted sources like the Leeds City Council's records to find the weather records after 2020. By integrating this external weather data into the cow growth dataset, future studies could more accurately assess the influence of these environmental factors on the growth of cows, leading to more coverage of the analysis of external conditions like weather for the farm animals.

Thirdly, this study was also constrained by computational limitations, which restricted the number of cross-validation iterations (as the tuning of AIC in lamb models is limited to only 20 times of cross-validation for the LMM of lambs, which has 4095 models), which could reduce the robustness of the model selection process. To mitigate this limitation, future research could utilize more advanced computational resources, such as higher CPU performance that allows for extensive model testing and validation through the CV process.

Moreover, the models' specificity to the conditions at Crag House Farm in Leeds may limit their applicability to adaptation to other farm settings. To overcome this problem, it is recommended that future studies be conducted in varied farming environments. For example, conducting research on different farms that might have different types of animals, such as pigs or different breeds of lambs and cows, could provide valuable insights and more coverage of longitudinal models on the other types of farm animals. Additionally, this approach could include evaluating different nutrition-focused management practices, such as bottle-feeding versus grassing, to validate the findings of different growth rates from the different styles of feeding. These suggestions could help ensure that the models are robust and applicable to other farm conditions.

By addressing these limitations through enhanced data collection, handling the uncovering of environmental considerations, increased computational power, and providing more validation in other farm settings, future research could build on this study's findings and limitations to develop more accurate and universally applicable models for predicting growth in farm animals.

6.2 Conclusion

This study has successfully addressed the two main objectives of this project, providing valuable insights into the factors influencing the growth and weight of farm animals and developing the model to predict the weight of farm animals, particularly lambs and cows, at Crag House Farm in Leeds. This longitudinal study utilises the application of LMMs that could handle the effect of the longitudinal structure of data, which repeated records of the same animal throughout time. The analysis from LMMs could be used to identify critical variables such as age, weight, genetic factors from breeders, and weather conditions that significantly impact animal growth, thereby fulfilling the first objective of understanding the factors affecting the animals' growth. Some of the important insights that could be received from the analysis are that older animals, higher weaning weight of animals, and selecting the breeder (specific sires) were found to have a statistically significant effect on greater weight gains for farm animals. In contrast, some factors, like birth type or the age of the breeder (mother), do not have a significant effect on the weight of farm animals.

In terms of the second objective, this objective aimed to tune the LMM to develop robust and accurate models and reduce the overfitting of the LMM to improve the performance of LMM to predict the weight of farm animals. The study successfully fine-tuned the LMM, enhancing its predictive accuracy and robustness by providing the statistically improved RMSE in both lamb and cow models, which is crucial for providing reliable recommendations to farmers.

However, this study also highlighted some limitations, such as the quality and completeness of the datasets and computational performance constraints (CPU performance) that limited the cross-validation time to tune the model and reduced the robustness and reliability of the models. Moreover, the limitations in this study also have an impact on the model applicability because the models developed in this study were effective for the specific datasets for Crag House Farm, but their applicability might be constrained and provide non-accuracy results when applied to other datasets, different types of animals, or other farm environments.

In terms of addressing the issues from limitations, future research should focus on improving data quality and completeness, utilising more powerful computational resources, and exploring more efficient algorithms to enhance model robustness. Additionally, future studies should aim to make the models more flexible to help the model more easily adapt to different animal species and environments. Moreover, the collaboration of other researchers could also help validate and expand the model's applicability across various farming scenarios.

In conclusion, this research has effectively answered the project's two main objectives, demonstrating the importance of data-driven decision-making in farm management. By leveraging longitudinal data and a statistical model, which is LMM, the study has provided a strong foundation for future research and practical applications in improving the growth and profitability of livestock.

Appendix A

Appendix: R Coding

A.1 R code for Plotting Weaning Weight of Lambs

```
# Stat test of lambs
library(ggplot2)

# Create the box plot 1
ggplot(day_1_weights, aes(x = factor(Year), y = Weight, fill = Sex)) +
  geom_boxplot() +
  facet_wrap(~ BirthType) +
  labs(title = "(a) Box Plots of Lamb Weaning Weight by Various Sex and
  Birth Type on Different Years",
       x = "Year",
       y = "Weaning Weight (kg)") +
  theme_minimal() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 10))

# Create the box plot 2
ggplot(lamb_data_plot, aes(x = factor(Year), y = Weight, fill = Sex)) +
  geom_boxplot() +
  labs(title = "(a) Box Plot of Weaning Weight by Sex Across Different
  Years",
       x = "Year",
       y = "Weaning Weight (kg)") +
  theme_minimal() +
  theme(legend.title = element_text(size = 12),
        legend.text = element_text(size = 12))
```

```
ggplot(lamb_data_plot, aes(x = factor(Year), y = Weight, fill = BirthType))
+
  geom_boxplot() +
  labs(title = "(b) Box Plot of Weaning Weight by Birth Type Across
  Different Years",
       x = "Year",
       y = "Weaning Weight (kg)") +
  scale_fill_manual(values = c("Single" = "#FFA500",
                               "Twin" = "#9ACD32",
                               "Triplet" = "#EE82EE",
                               "Quadruplet" = "#F0E68C")) +
  theme_minimal() +
  theme(legend.title = element_text(size = 12),
        legend.text = element_text(size = 12))
```

A.2 R code for Plotting Weaning Weight of Cows

```
# Box plot to compare CalfWeight across different BornYear
ggplot(cow_unique, aes(x = factor(BornYear), y = CalfWeight, fill = Sex)) +
  geom_boxplot() +
  labs(title = "(a) Box Plot of Calf Weight on Weaning Date
  in Different Birth Years",
       x = "Birth Year",
       y = "Calf Weight on Weaning Date (kg)") +
  theme_minimal() +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        axis.title = element_text(size = 14),
        axis.text = element_text(size = 12),
        plot.title = element_text(size = 14, face = "bold"))

# Scatter plot of CalfWeight vs. MotherAge
ggplot(cow_unique, aes(x = MotherAge, y = CalfWeight)) +
  geom_point(alpha = 0.7, size = 3, color = "darkgreen") +
  labs(title = "(b) Relationship Between Maternal Age and
  Calf Weaning Weight",
       x = "Mother Age (years)",
       y = "Calf Weight on Weaning Date (kg)") +
  theme_minimal(base_size = 15) +
  theme(legend.position = "none",
        plot.title = element_text(size = 14, face = "bold"),
        plot.subtitle = element_text(size = 14),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 14),
        panel.grid.major = element_line(color = "grey80"),
        panel.grid.minor = element_blank())
```

A.3 Statistical Test to Compare the Weaning Weight of Lambs

```
# Filter the data to keep only male and female lambs and Weaning Weight
lamb_data_mf_day1 <- lamb_data %>%
  filter(Sex %in% c("M", "F") & Day == 1)

# independent t-test
t_test_result <- t.test(Weight ~ Sex, data = lamb_data_mf_day1)

print(t_test_result)

# pairwise t-tests
pairwise_results <- pairwise.t.test(lamb_data_unique$Weight,
    lamb_data_unique$BirthType)

print(pairwise_results)
```

A.4 Statistical Test to Compare the Weaning Weight of Cows

```
# Filter the data to keep only male and female calves
cow_mf <- cow_unique %>%
  filter(Sex %in% c("M", "F"))

# Independent t-test
t_test_result <- t.test(CalfWeight ~ Sex, data = cow_mf)

print(t_test_result)
```

A.5 EDA: Histogram Distribution of Lambs

```
# Select specific numerical columns
selected_columns <- lamb_data %>%
  select(DW, Weight, WeaningWeight, Age = AgeMonths, MotherAge,
  WindSpd, Temp8m, GlobRad, Humidity) %>%
  mutate(Age = Age / 12)

plot_list <- list()

# Define custom binwidths
binwidths <- c(DW = 1, Weight = 2, WeaningWeight = 2, Age = 0.5,
MotherAge = 1, WindSpd = 5, Temp8m = 1, GlobRad = 20, Humidity = 5)

# Create histograms
for (i in 1:ncol(selected_columns)) {
  column_name <- colnames(selected_columns)[i]
  binwidth <- binwidths[[names(binwidths)[i]]]

  # Rename columns for better labels
  if (column_name == "DW") column_name <- "Dead Weight (kg)"
  if (column_name == "Weight") column_name <- "Weight (kg)"
  if (column_name == "WeaningWeight") column_name <- "Weaning Weight (kg)"
  if (column_name == "Age") column_name <- "Age of Lambs (year)"
  if (column_name == "MotherAge") column_name <- "Mother Age (year)"
  if (column_name == "WindSpd") column_name <- "Wind Speed (m/s)"
  if (column_name == "Temp8m") column_name <- "Temperature (degree Celcius)"
  if (column_name == "GlobRad") column_name <- "Global Radiation
  (Watt Per Square Meter)"
  if (column_name == "Humidity") column_name <- "Humidity (%)"

  p <- ggplot(selected_columns, aes_string(x = names(binwidths)[i])) +
    geom_histogram(binwidth = binwidth, fill = "#69b3a2",
    color = "#1f78b4", alpha = 0.7) +
    labs(title = column_name, x = "Value", y = "Count") +
    theme_minimal() +
    theme(
      plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
      axis.title.x = element_text(size = 12),
      axis.title.y = element_text(size = 12),
```

```
      axis.text.x = element_text(size = 10),
      axis.text.y = element_text(size = 10)
    )
  plot_list[[i]] <- p
}

# Combined plot with a main title
grid.arrange(
  arrangeGrob(
    grobs = plot_list,
    ncol = 3,
    top = textGrob("Distribution of Numerical Columns in the
    Combined Lamb Dataset", gp = gpar(fontsize = 20, fontface = "bold"))
  )
)
```

A.6 EDA: Growth Trends of Lambs

```
# Filter the data for the year 2017 and exclude Day 1
lamb_data_2017 <- lamb_data %>% filter(Year == 2017 & Day > 1)

# Different Type
ggplot(lamb_data_2017, aes(x = Day, y = Weight, color = type, group = LambID)
  geom_line(size = 0.7) +
  geom_point(size = 2) +
  scale_color_brewer(palette = "Dark2") +
  labs(title = "(b) Weight Growth of Lambs by Type over Days in 2017",
       x = "Day", y = "Weight (kg)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )
# Different Sex
ggplot(lamb_data_2017, aes(x = Day, y = Weight, color = Sex, group = LambID))
  geom_line(size = 0.7) +
  geom_point(size = 2) +
  labs(title = "(a) Weight Growth of Lambs by Sex over Days in 2017",
       x = "Day", y = "Weight (kg)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )
```

A.7 EDA: Correlation Plot of Lambs

```
install.packages("corrplot")
library(corrplot)

# Select numerical columns
selected_columns <- lamb_data %>% select(DW, Weight, AgeMonths,
WeaningWeight, MotherAge, WindSpd, Temp8m, GlobRad, Humidity) %>%
  rename(
    `Dead Weight` = DW,
    `Weight` = Weight,
    `Age of Lambs` = AgeMonths,
    `Weaning Weight` = WeaningWeight,
    `Mother Age` = MotherAge,
    `Wind Speed` = WindSpd,
    `Temperature` = Temp8m,
    `Global Radiation` = GlobRad,
    `Humidity` = Humidity
  )

# Calculate the correlation matrix
correlation_matrix <- cor(selected_columns, use = "complete.obs")

# Plot the correlation matrix
corrplot(correlation_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black",
         title = "Correlation Plot of Selected Numerical Columns in the
         Lamb Dataset",
         mar = c(0, 0, 1, 0))
```

A.8 Fitting the LMM of Lambs

```
# Install and load lm4
install.packages("lme4")
library(lme4)

# selected the columns that use for analysis
selected_data <- lamb_data %>%
  select(ID = No., Age = AgeMonths, Weight, WeaningWeight, Sex,
  Type = type, SireID = Sire, AgeOfMother = MotherAge, BirthType,
  WeightSeason, WindSpeed = WindSpd, Temperature = Temp8m,
  GlobalRadiation = GlobRad, Humidity) %>%
  mutate(Age = Age / 12) %>%
  mutate_if(is.numeric, ~ round(., 3))

# Fitting the LMM of lambs
model <- lmer(Weight ~ Age + Sex + Type + WeaningWeight + SireID +
    AgeOfMother + BirthType + WeightSeason + WindSpeed + Temperature +
    GlobalRadiation + Humidity + (1 | ID), data = selected_data)

# Summaries model
summary(model)

# Calculate 95% confidence intervals for the fixed effects
conf_intervals <- confint(model, level = 0.95)

print(conf_intervals)
```

A.9 Fitting the LMM of Cows

```
# Fit the LMM of Cows
library(lme4)
library(dplyr)

# Select the columns that use for the analysis
selected_data <- cow %>%
    select(CalfID, Age = AgeMonths, Weight, CalfWeight, Sex, SireName=
    Sire, AgeOfMother = MotherAge, WeightAtSeason) %>%
  mutate(Age = Age / 12) %>%
  mutate_if(is.numeric, ~ round(., 3))

# Fit the mixed-effects model
model <- lmer(Weight ~ Sex + WeightAtSeason + CalfWeight + SireName +
WeightAtSeason + AgeOfMother + Age + (1 | CalfID), data = selected_data)

summary(model)

# Calculate 95% confidence intervals for the fixed effects
conf_intervals <- confint(model, level = 0.95)

print(conf_intervals)
```

Bibliography

[1] Arnold, K., Subotnik, R., and Ross, M. 2011. *Encyclopedia of Creativity. Longitudinal Studies*. 2nd ed. Amsterdam: Academic Press.

[2] Assan, N. and Makuza, S. 2005. The Effect of Non-genetic Factors on Birth Weight and Weaning Weight in Three Sheep Breeds of Zimbabwe Anim Biosci. *Asian-Australasian Journal of Animal Sciences*. **18**(2), pp.151-157.

[3] Bailey, J. 2008. First steps in qualitative data analysis: Transcribing. *Family Practice*. **25**(2), pp. 127-131.

[4] Coleman, L., Back, P., Blair, H., and Hickson, R. 2021. Sire Effects on Birth Weight, Gestation Length, and Pre-Weaning Growth of Beef-Cross-Dairy Calves: A Case Study in New Zealand. *Dairy*. **2**(3), pp. 385-395.

[5] Crawshaw, H. 2023. Guide to British cattle breeds: common cow breeds and how to recognise them. 17 May. *BBC Countryfile*. [Online]. [Accessed 3 August 2024]. Available from: https://www.countryfile.com/wildlife/mammals/native-british-cattle-breeds-and-how-to-recognise-themhereford

[6] De Rose, E.P. and Wilton, J.W. 1991. Productivity and profitability of twin births in beef cattle. *Journal of Animal Science*. **69**(8), pp. 3085-3093.

[7] Elbert, K., Matthews, N., Wassmuth, R., and Tetens, J. 2020. Effects of sire line, birth weight and sex on growth performance and carcass traits of crossbred pigs under standardized environmental conditions, *Arch Anim Breed*. **63**(2), pp. 367–376.

[8] Fisher, M.W. 2007. Shelter and welfare of pastoral animals in New Zealand. *New Zealand Journal of Agricultural Research*. **50**(3), pp. 347–359.

[9] Gaskins, C.T., Snowder, G.D., Westman, M.K., and Evans, M. 2005. Influence of body weight, age, and weight gain on fertility and prolificacy in four breeds of ewe lambs. *Journal of Animal Science*. **83**(7), pp. 1680-1689.

[10] Hasan, M.A., Hasan, M.K., and Mottalib, M.A. 2015. Linear regression-based feature selection for microarray data classification. *International Journal of Data Mining and Bioinformatics*, **11**(2), pp. 167-179.

[11] Hatcher S., Eppleston J., Graham R.P., McDonald J., Schlunke S., Watt B. and Thornberry K.J. 2008. Higher weaning weight improves post-weaning growth and survival in young Merino sheep. *Australian Journal of Experimental Agriculture*. **48**, pp. 966-973.

[12] Jennings, G.R. 2005. *Encyclopedia of Social Measurement*. Business, Social Science Methods Used in. 1st ed. Amsterdam: Elsevier.

[13] Jumah, F., Raju, B., Nagaraj, A., Shinde, R., Lescott, C., Sun, H., Gupta, G., and Nanda, A. 2022. Uncharted Waters of Machine and Deep Learning for Surgical Phase Recognition in Neurosurgery. *World Neurosurgery*. **160**, pp. 4-12.

[14] Karabulut, E.M., Özel, S.A., and İbrikçi, T. 2012. A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*. **1**, pp. 323-327.

[15] Leeds City Council. 2020. *Leeds meteorological data*. [Online]. [Accessed 16 July 2024]. Available from: https://www.data.gov.uk/dataset/4afd0747-4fba-49c4-b0aa-b6f093a7db2c/leeds-meteorological-data

[16] Leitner, W. and Turner, W.R. 2017. *Encyclopedia of Biodiversity*. 2nd ed. Levin, S.A. ed. Academic Press.

[17] Liamputtong, P. 2009. Qualitative data analysis: Conceptual and practical considerations. *Health Promotion Journal of Australia*. **20**(2), pp.133-139.

[18] Liu, X. 2015. *Methods and Applications of Longitudinal Data Analysis*. 1st ed. Academic Press.

[19] Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., and Forghani, R. 2020. Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment. *Neuroimaging Clinics of North America*. **30**(4), pp. 433-445.

[20] Met Office. 2024. When does spring start?. *Met Office Weather*. [Online]. [Accessed 3 July 2024]. Available from: https://www.metoffice.gov.uk/weather/learn-about/weather/seasons/spring/when-does-spring-start

[21] Mohammed, E.A., Naugler, C., and Far, B.H. 2015. Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*, pp. 577-602.

[22] Murphy, J.I., Weaver, N.E., and Hendricks, A.E. 2022. Accessible analysis of longitudinal data with linear mixed effects models. *Disease Models and Mechanisms*. **15**(5), [no pagination].

[23] Neill, S.P., and Hashemi, M.R. eds. 2018. *Fundamentals of Ocean Renewable Energy*. Academic Press.

[24] Olanrewaju, A.T., Hossain, M.A., Whiteside, N., and Mercieca, P. 2020. Social media and entrepreneurship research: A literature review. *International Journal of Information Management*. **50**, pp. 90-110.

[25] Pusponegoro, N.H., Rachmawati, R.N., Notodiputro, K.A., and Sartono, B. 2017. Linear Mixed Model for Analyzing Longitudinal Data: A Simulation Study of Children Growth Differences. *Procedia Computer Science*. **116**, pp. 284-291.

[26] Rao, C.R., Miller, J.P. and Rao, D.C. eds. 2011. *Essential Statistical Methods for Medical Statistics*. North-Holland.

[27] Saraux, C., Chiaradia, A., Salton, M., Dann, P. and Viblanc, V.A., 2016. Negative effects of wind speed on individual foraging performance and breeding success in little penguins. *Ecological monographs*. **86**(1), pp.61-77.

[28] Schneider, P. and Xhafa, F. eds. 2022. Anomaly Detection and Complex Event Processing over IoT Data Streams. Academic Press.

[29] Simeonov, M., Todorov, N., Nedelkov, K., Kirilov, A. and Harmon, D.L., 2014. Influence of live weight, sex and type of birth on growth and slaughter characteristics in early weaned lambs. *Small Ruminant Research*. **121**(2), pp.188-192.

[30] Smith, S.P., Pollak, E.J., and Anderson, G.B. 1982. Maternal Influences on Growth of Twin and Single Calves. *Journal of Animal Science*. **55**(3), pp. 533-542.

[31] Sugiyama, M. ed. 2016. *Introduction to Statistical Machine Learning*. Morgan Kaufmann.

[32] Tiwari, K. and Chong, N.Y. 2020. *Multi-robot exploration for environmental monitoring: The resource constrained perspective*. Cambridge: Academic Press.

[33] Toga, A.W. ed. 2015. *Brain Mapping: An Encyclopedic Reference*. Academic Press.

[34] Pandey, R., Khatri, S.K., Singh, N.K., and Verma, P. eds. 2022. *Artificial Intelligence and Machine Learning for EDGE Computing*. 1st ed. London: Academic Press.