



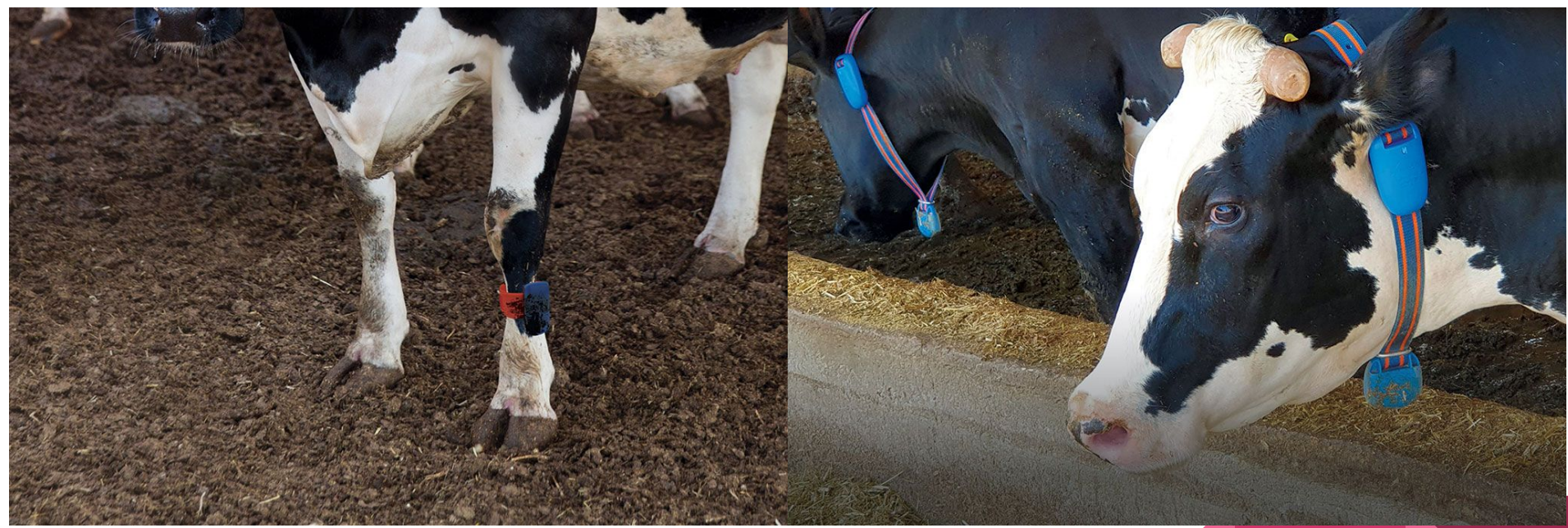
Bovine Event Detection: Analysis of Bovine Data and Cycles

By: Emily Medema
Supervisor: Dr. Ramon Lawrence



Motivation

- The agricultural industry, specifically bovine, is a prominent industry in Canada with incredible potential for data analytics.
- Current data is either proprietary or stored in spreadsheets on personal computers and not shared.
- Filling this gap of open-sourced data analysis will allow for a more customized analysis to be performed on bovine data, widening the potential research avenues utilizing sensor data.



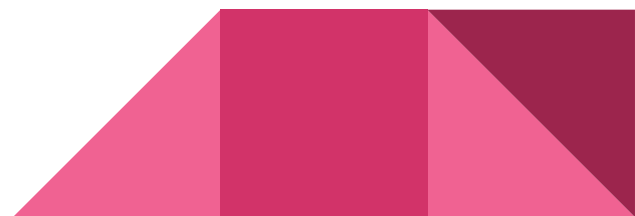
Overview - Plan for Bovine Data

As a whole, our goal was to process data stored in a collection of spreadsheets and transferring it to a MySQL database.

After which, we created a quick site that allows us to take a look at the data.

We then created models that will allow us to detect outliers within the data.

Due to the success of outlier detection, we then moved on to Event Detection.



Spreadsheet Conversion

The spreadsheet conversion is done through one python script.

The script does the following:

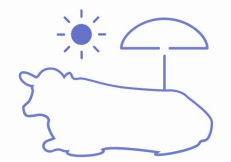
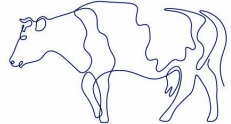
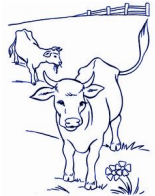

1. Connects to the database through the connection information specified in the `custom.py` file.
2. Reads in the first spreadsheet found in the `datasets` folder.
3. Uploads the data to the corresponding table within the database.
4. Moves file from the `datasets` folder to `processed` folder.
5. Repeats step 2 through 4 until there are no more files in `datasets`.

Data Cleaning

In order to receive meaningful results, data cleaning is critical.

This processing involves reducing data noise, imputing missing data, outlier detection, and data aggregation.



Table	Columns	Explanation
AfiAct2 Rest Hourly 	<ul style="list-style-type: none"> • Hour • RestTime • RestBout 	<ul style="list-style-type: none"> • Impute any missing hours of the day • Remove outliers and impute missing values
AfiAct2 Steps and Activity 15 minutes 	<ul style="list-style-type: none"> • Step • DateTime_Collected 	<ul style="list-style-type: none"> • Derive column for step difference • Impute any missing 15 minutes of the day
AfiCollar Motion Hourly 	<ul style="list-style-type: none"> • Hour • MotionHeatIndicator • Motion 	<ul style="list-style-type: none"> • Impute any missing hours of the day • Remove outliers and impute missing values
AfiCollar Rumination and Eating Hourly 	<ul style="list-style-type: none"> • Hour • RuminationTimeInSeconds • EatingTimeInSeconds 	<ul style="list-style-type: none"> • Impute any missing hours of the day • Remove outliers and impute missing values

Outlier Detection

To demonstrate some of the potential we have for data analysis now that we have our data in the database, we analyzed the data to see if we can determine when the data is an outlier.

With this we will be able to tell if the behaviour of a cow is odd or in statistical terms an outlier.

We have done this in a few different ways.



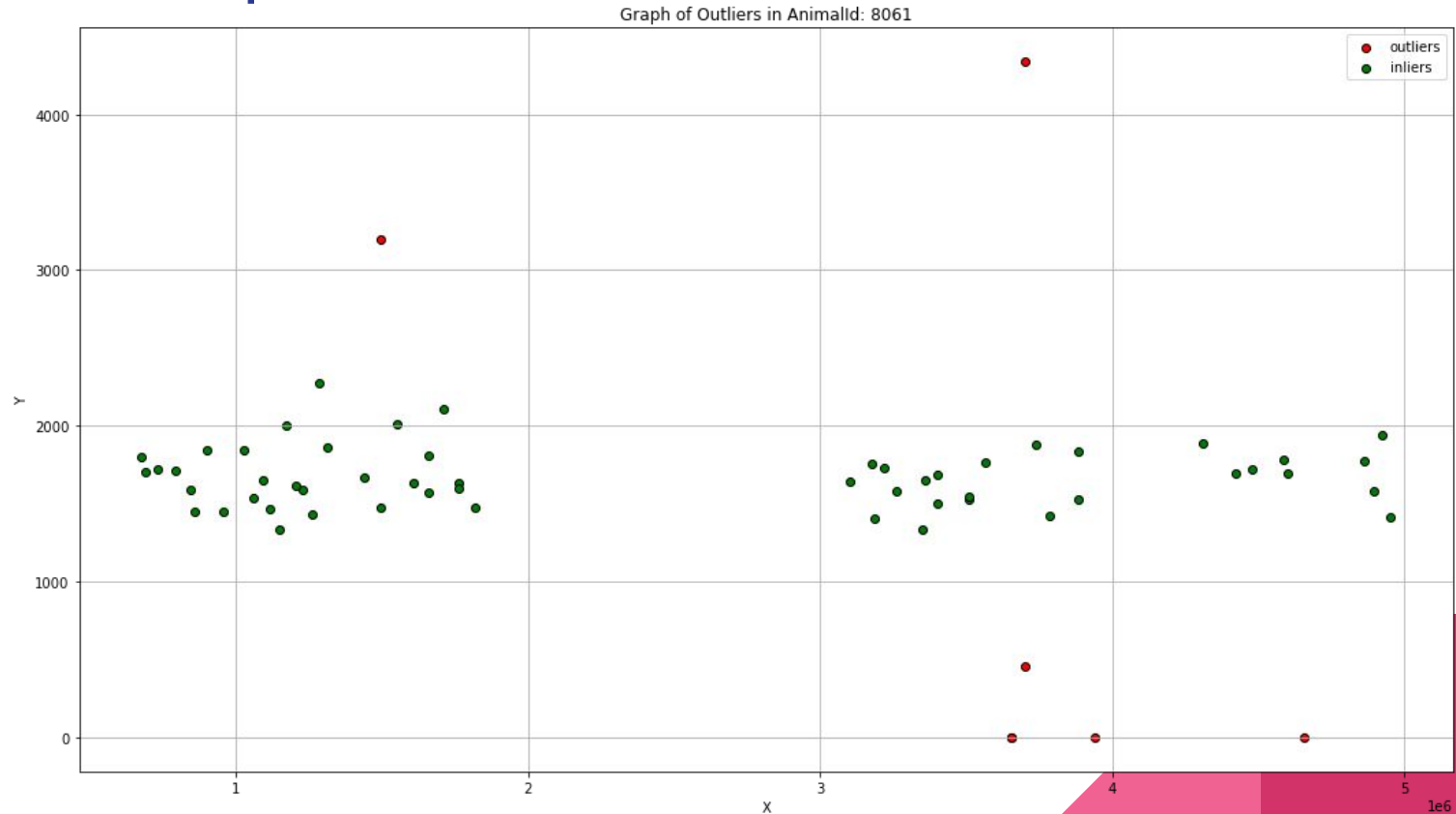
Outlier Detection - KNN

K Nearest Neighbours (KNN) is a supervised learning technique that focuses on classification.

This model forms neighbourhoods based on a measure of closeness. If a value is too far it can be considered an outlier.

This model works for what we want to do, but we cannot extend it properly to multiple variables and as a supervised learning technique it is not our optimal choice.

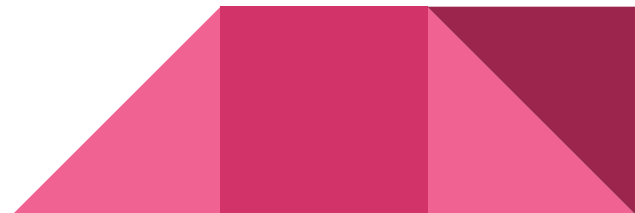
KNN on Steps of Cow 8061



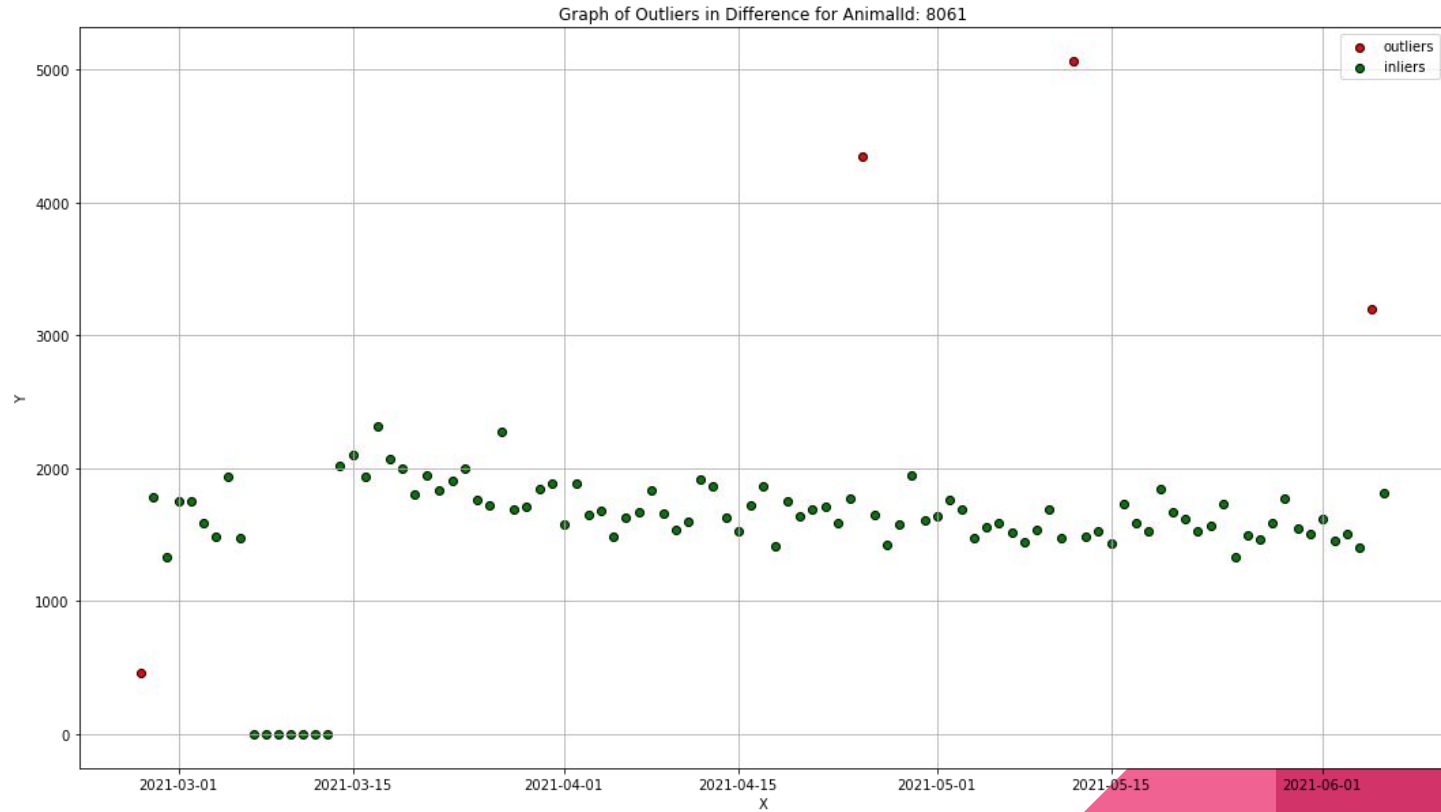
Outlier Detection - Isolation Forest

Isolation Forests' are a unsupervised learning technique used for classification and outlier detection.

Extension of the popular random forest algorithm. The building blocks of isolation forests are isolation trees with a binary outcome (is/is not an outlier).



Isolation Forest - Steps for Cow 8061



Outliers in the Data

We can look at the commonly marked outliers by our models and use them to see if there is any correlation within the tables.

We can see the common outliers here:

outliers_if				
	Date_Collected	variable_sum	ids	variable_sum_Isolation_Forest_Anomaly
0	2021-02-26	460.0	3941924	True
58	2021-04-25	4340.0	4658861	True
75	2021-05-12	5062.0	3317191	True
99	2021-06-05	3199.0	1494339	True

outliers_knn			
	Date_Collected	variable_sum	ids
0	2021-02-26	460.0	3941924
9	2021-03-07	0.0	3700972
10	2021-03-08	0.0	3700430
11	2021-03-09	0.0	3656866
12	2021-03-10	0.0	3656418
58	2021-04-25	4340.0	4658861
99	2021-06-05	3199.0	1494339

Event Detection

In our attempt to detect outliers, we unintentionally created an event detection model, specifically an *estrus* event detection model.

Estrus is a state within the estrous cycle within a cow that precedes ovulation and only lasts between 6 and 20 hours every ~21 days.

This is the main focus of industry as *estrus* detection is important for the reproductive management of the farm.

Current Event Detection

Estrus detection is either through the company providing the sensors or through the calculation of rolling averages and other measures.

In prior research this event detection was done retrospectively from the data in the csv files and through the coding language R.

R Translation

- As we already have a model that is seemingly detecting *estrus*, we also wanted to have a comparison of the current detection within the same language as our machine learning model.
- We also wanted to ensure that there be as little of a learning curve as possible for future workers.
- Python also has access to different libraries and models that might be helpful.

Challenges of R Translation

- Library mismatches
- Different interactions with dataframes
- Different functions

Results

- A script to upload data from a CSV file to a MySQL database is working and customizable
- The temporary MySQL database can be accessed remotely and is hosted securely on a Digital Ocean server
- Data overview and cleaning is completed
- The developed isolation forest model detects outliers and depending on the variable used *estrus*
- The previous event detection code has been translated from R to Python.

Results - R Translation

There were approximately 173 events detected via the R code, however, the Python code detected around 500.

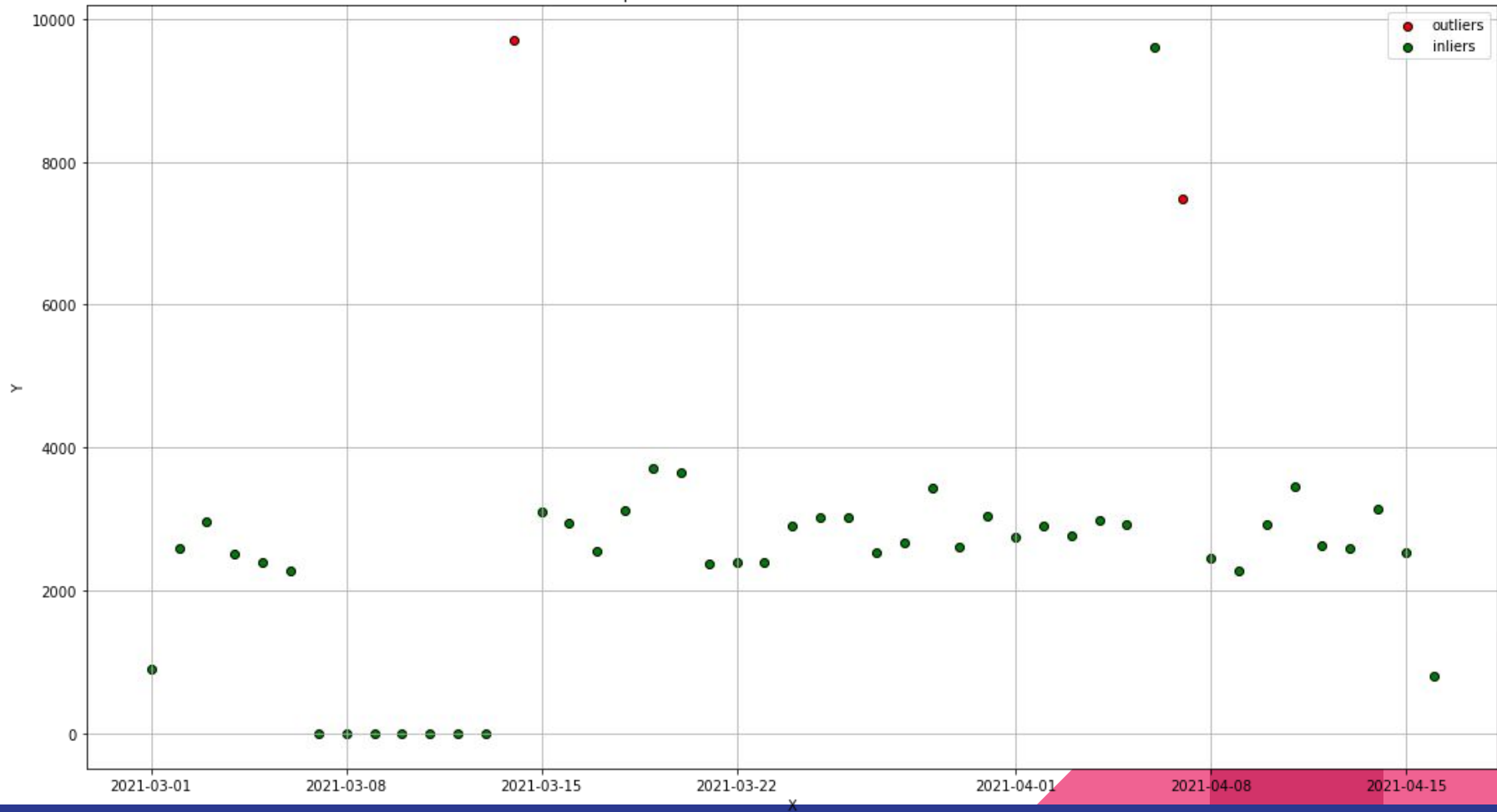
These false positives and negatives are most likely due to the rolling mean and Kalman Filter imputation calculations as well as the thresholds.

Results - Isolation Forest vs. Rolling Average

The Isolation Forest has less false positives than the translation of the code that is currently used for *estrus* detection.

When the Isolation Forest model is run on Animal 66, we can see that it detect "outliers" of a high amount of steps in a potential 24 day cycle.

Graph of Outliers in Difference for AnimalId: 66



Results - Rolling Average vs. Isolation Forest

```
df_step15[df_step15.id == 66]
```

	id	group_est	start_est	last_est	dur5
304	66	610	2021-03-11 07:00:00	2021-03-11 10:00:00	4
305	66	612	2021-03-13 17:00:00	2021-03-14 02:00:00	14
306	66	614	2021-03-14 04:00:00	2021-03-14 11:00:00	8
307	66	616	2021-03-20 08:00:00	2021-03-20 11:00:00	4
308	66	618	2021-03-29 07:00:00	2021-03-29 12:00:00	6
309	66	620	2021-03-31 07:00:00	2021-03-31 11:00:00	5
310	66	622	2021-04-01 08:00:00	2021-04-01 11:00:00	4
311	66	624	2021-04-06 05:00:00	2021-04-07 01:00:00	3
312	66	626	2021-04-07 06:00:00	2021-04-07 12:00:00	7
313	66	628	2021-04-11 08:00:00	2021-04-11 11:00:00	4

```
outliers_if
```

	Date_Collected	variable_sum	ids	variable_sum_Isolation_Forest_Anomaly
13	2021-03-14	9700.0	3539685	True
37	2021-04-07	7473.0	4967381	True

Results - R Rolling Average vs. Isolation Forest

id	fract	date_time1	est	duration1	intensity_mean	intensity_max	group_est	start_est	last_est	dur5
66	2	2021-03-13 23:00:00	EST	9	520.8746	683.4722	2	2021-03-13 23:00:00	2021-03-14 07:00:00	9
66	0	2021-04-06 05:00:00	EST	17	328.6283	587.2881	4	2021-04-06 05:00:00	2021-04-06 21:00:00	17

outliers_if

	Date_Collected	variable_sum	ids	variable_sum_Isolation_Forest_Anomaly
13	2021-03-14	9700.0	3539685	True
37	2021-04-07	7473.0	4967381	True

Conclusion

Now that the data is accessible to researchers and farmers, the data can be used in a variety of ways.

With the input of researchers and farmers, the detection models and analysis can become better and more useful than ever.



Future Works

It is clear that there is potential for a multitude of different projects and paths within this field:

- There are some issues with the R translation/event detection through the rolling average and thresholds that must be resolved.
- An expansion of the data through external validation and feature engineering as well as classification of the animals would be a great next step toward creating an amalgamation of all the animals and their physiology and genetic classifications.



Questions?