# Group ID - MSc in Data Analytics

Author: Rakesh Kumar Muraleedharan

e-mail: [2022589@student.cct.ie](mailto:2022589@student.cct.ie)

Student ID: 2022589

**Organic Livestock Analysis and Prediction**

**Goal:** To predict the organic livestock count of Ireland, identify the animal which shows highest count and then compare with other countries in Europe.

***ABSTRACT:***

*Organic farming can be defined as a method of production which places the highest emphasis on environmental protection and, regarding livestock production, on animal welfare considerations. It avoids or largely reduces the use of synthetic chemical inputs such as fertilisers, pesticides, additives and medicinal products. The production of genetically modified organisms (GMOs) and their use in animal feed are forbidden. It is as a part of a sustainable farming system and a viable alternative to the more traditional approaches to agriculture.*

*EU-aggregates are calculated as far as national data are available for all Member States.*

*From 2012 data onwards*

*For our analysis here we will just be focussing on the Organic livestock data.*

*EDA is performed on the data set, identified the gaps in the data, bad data cleaned up using appropriate methods. Descriptive statistics are done to identify maximum /minimum/average counts are conducted. Appropriate charts are drawn. Inferential Statistics is performed on the livestock count with a specific confidence interval. Hypothesis testing is carried out for single population Ireland as well as comparing with multiple populations for Ireland and other EU countries (Romania/Greece/Hungary) taken as an example.*

*Since the data available is continuous numerical, linear regression and regression trees were applied. The training accuracy was extremely high, but testing was incredibly low, leading to overfitting. This necessitated the need to do data enrichment and made use of additional data clean up and aggregation*

*All the above steps were done in combination. Cross-validations were done against the predicted data and the trained dataset.*

*One of the important activity parts of the research is continuous data processing, this is done in iteration depending on how best the machine model learning reacts.*

*The most accurate forecast was seen in Linear Regression with GRID CV.*

*Sentimental analysis was done based on the livestock counts greater than or less than their respective averages.*

**Keywords:**

Data Analytics, Data Preparation, Statistical Analysis, Machine Learning, PYTHON, Regression Models, Visualisation, Data Pre-processing, Dublin Footfall data.

1. **Introduction:**

Farming is organic if it complies with 'Council Regulation (EC) No 834/2007 of 28 June 2007 on organic production and labelling of organic products and repealing Regulation (EEC) No 2092/91.Organic farming differs from other agricultural production methods in the application of regulated standards (production rules), compulsory control schemes and a specific labelling scheme.

This study aims to analyse and predict the organic livestock count for Ireland, compare with other European countries. This study might be useful for governments to consider expanding their investments in these livestock.

1. **Methodology:**

The project-management methodology adopted in this research is a combination of KDD (Knowledge Discovery is Database) and CRISP-DM (Cross-Industry Standard Process for Data Mining).

The activities involved are research understanding, data collection, data preparation, Data Analysis, Statistical Analysis, enrichment by collecting additional data, and discussion of analysis results. The method of predictive analysis, based on machine learning was chosen. Training, validation, and testing of machine learning models were performed. A complete set of activities is performed programmatically using the open-source software PYTHON.

Below is how the entire analysis is designed.

* 1. **Research Understanding Phase:**

Understanding of Irish agriculture, understanding the topology and its impact on the livestock. Understanding the impact of organic farming on the environment.

The next step is to define the scope of analysis, this should align with the project guidelines (CA2 brief provided by the college)

This is then followed by browsing different websites to get authentic reliable data.

* 1. **Data Collection :**

The data required for this project is obtained from the European Union Website: <https://ec.europa.eu/eurostat/databrowser/view/ORG_LSTSPEC/default/table?lang=en&category=agr.org>.

The data collection was based on the Commission Regulation (EC) No 889/2008 implementing the Council Regulation (EC) No 834/2007 on organic production and labelling of organic products. This data was collected as summary tables to display an overview of the uptake of organic farming within the European Union, the United Kingdom, Iceland and Norway. Switzerland and some candidate countries (Montenegro, North Macedonia, Serbia and Turkey) also provide data on a voluntary basis.

In some cases, there were zero counts shown up, this doesn’t really mean there isn’t any livestock in the country for that year, but it means the numbers are not provided by the member country.

The data is available as a open data in the European Union website, confidentiality of the data is described as a column marked as ‘c’, this is not applicable in the data selected for organic livestock analysis.

Metadata for the acquired data is available below:

<https://ec.europa.eu/eurostat/cache/metadata/en/org_esms.htm>

The data is available in a data browser, it is downloaded and used for analysis.

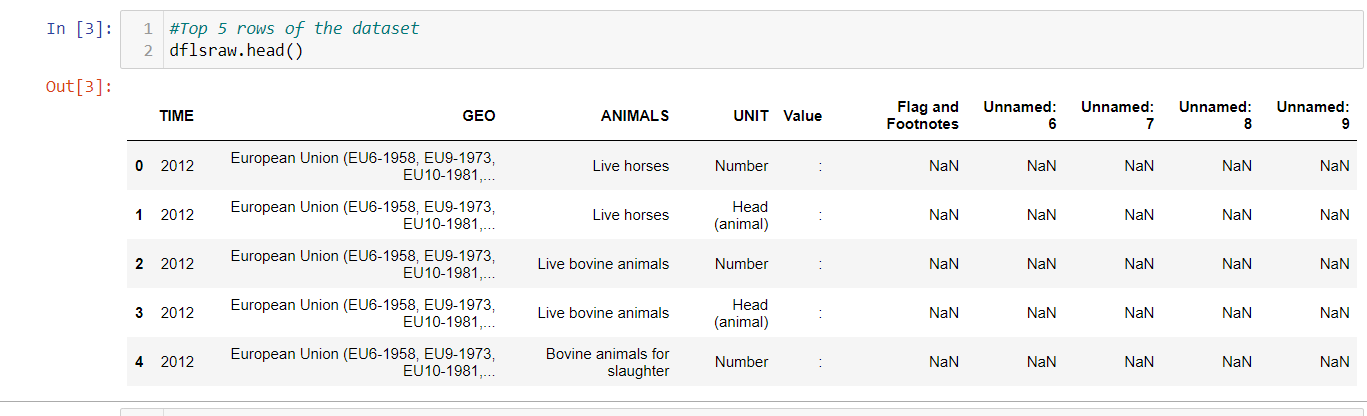
* 1. **Data Preparation**

In the first part of this research, data preparation was performed to obtain data suitable for analysis and apply appropriate machine learning models.

1. The dataset contained several records without any meaningful statistics.

Example: The summarized data for European Union, the rows corresponding were not complete across different years, and it didn’t serve the purpose of analysing Ireland specific data and comparison with few specific countries, hence such records were removed, blank columns removed.

1. There are two units Number and Head, we will keep one UNIT and base our analysis on this (Head).



1. The column indicated the measure to be analysed/predicted is a object, and it contains characters (,) data not provided was marked as ‘:’ this needs to be removed.
2. UK is not part of EU, hence removing them from the dataset.
3. After data clean up, EDA is done to identify the NULL values,
4. Further operations on data, to reorganise the rows/columns using pivot, aggregate the data, so that data is ready for doing the Inferential statistics/ Hypothesis testing.
5. Data is further granulated to do regression analysis, a new field is created called sentiment, sentiment analysis is performed on this column.
   1. **Modelling Phase**:

In this research, we have used methods of supervised machine learning. Building the machine learning model consisted of the following phases:

1. Defining the goal of the model.
2. Choosing dependent variables i.e., the dataset attribute which value we want to predict using the machine learning model. The footfall counts were selected as the target.
3. Select relevant attributes (features) of a dataset. Corelation needs to be established to select the correct attributes for the right prediction.
4. Selecting a supervised machine learning algorithm, according to the nature of labels and attributes. As the target variables were numerical, I used regression algorithms such as linear regression, and regression trees.
5. Datasets (training and test) pre-processing that fulfil requirements of the selected algorithm.
6. Model tuning – setting hyperparameters that are specific for each type of machine learning algorithm.
7. Model training – application of the selected machine learning algorithm on the training dataset to obtain model parameters.

**2.5 Model Evaluation:**

1. Model evaluation using cross-validation. Cross-validation is a method for getting a reliable estimate of model performance using only training data.
2. Model Testing – to predict the performance of a model on a new dataset, we need to assess its performance measures on a dataset that played no part in the formation of the model. This independent dataset is called the test dataset. Comparing test vs training performance allows us to avoid overfitting. If the model performs very well on the training data but poorly on test data, then it is overfitting.
3. Selecting a winning model – a model that has the best performance on the test dataset.
4. Label prediction using the winning model.
5. As per the requirement for the CA, there is an exclusive sentiment analysis done based on the livestock count above or below the average.

**2.6** **Deployment Phase:**

Once all the Machine Learning models, best model is deployed to live for future predictions.

1. **Detailed Analytical Steps:**

The detailed steps described here are inline with the CRISP model explained above.

* 1. **Programming Language:**

PYTHON is the programming language used in this research, starting from data acquisition, data enrichment, data preparation, and visualization.

The coding is done using PYTHON, including the usage of conditional statements *(if loop),* iterations *(for loop)*, and classes/objects, lambda functions etc.

PYTHON libraries are used including Pandas (data manipulation and analysis), NumPy - to work with arrays and numerical operations, Calendar for DateTime operations, matplotlib and seaborn for visualization, sklearn for machine learning including model selection, pre-processing, model evaluation, validations, etc.

Statistical analysis including descriptive statistics, distributions, inferential statistics, Hypothetical testing are all developed using PYTHON libraries SciPy, charts are drawn using the matplot lib

The code is commented clearly on what each part does.

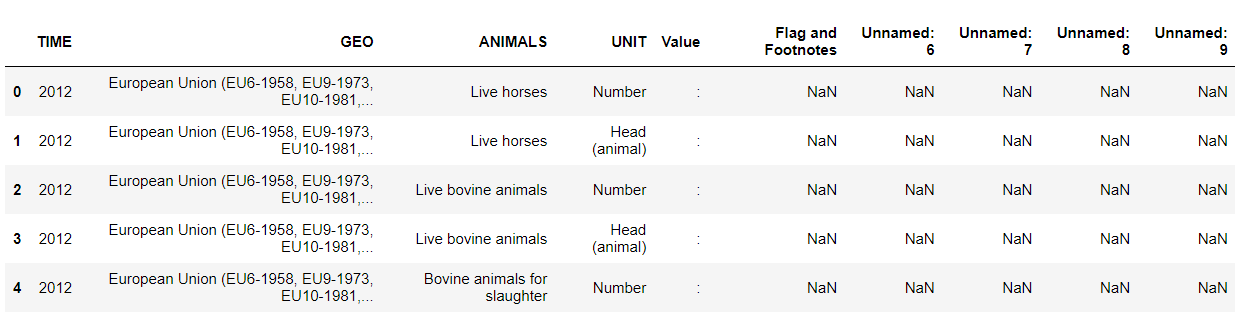
The code is developed in Jupyter notebook. It is an interactive development environment used most commonly for development using PYTHON, it will assist in the coding standard, indentation, and syntax of the code. The notebook is especially useful in-class assignments where the code can be shared and used.

* 1. **Exploratory Data Analysis:**

EDA is most critical in any data analytic project. EDA consists of data clean up and aggregation that would suit the final modelling requirements.

The source data for this analysis is from European Union website, this contains data for multiple countries, and an aggregated value.

Below is how the raw data looked like :



At the first glance of the data, we can see there are extra fields with NaN, they do not make any sense for our analysis here, hence we drop them.

A glance at the unique countries:



The aggregated values in the EU will lead to incorrect model selection, since it doesn’t need to be there for our analysis, we will drop them too.

UK is also removed from the dataset.

Below livestock has no data available, hence dropping them

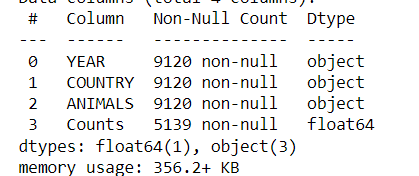
*Bees (hives)*

*Ewes and ewe-lambs put to the ram*

*Laying hens (producing eggs for consumption)*

Blank Values are denoted by ‘:’ , we will remove them appropriate based on the amount of such bad data. We will first convert ‘:’ to NaN to ease our analysis.

*Pic: Number of NULL values for the field counts*

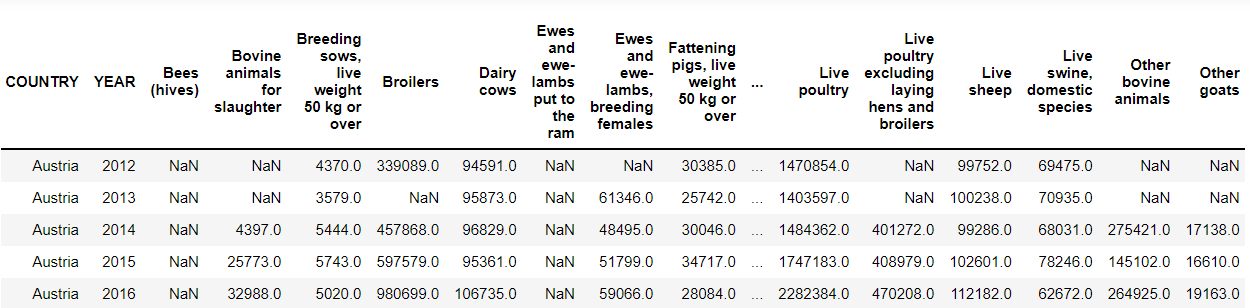




There are more than 50% values NULL, hence we cannot drop them.

Next step we will use the pivot function of PYTHON , and realign the data, so that each column represents the value for each livestock , and every row corresponds to each Year and country.

*Pic: Top 5 rows after pivot*



*Pic: Analysis showing the data variations for each animal*

**Table

Description automatically generated with low confidence**

Below livestock has no data available, hence dropping them

*Bees (hives)*

*Ewes and ewe-lambs put to the ram*

*Laying hens (producing eggs for consumption)*

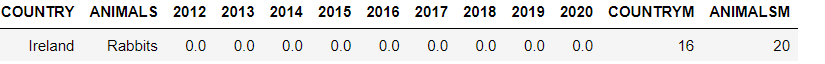
Next step, we will replace the NaN value in each of the Live Stock category, with the value in the preceding year, if there isnt any previous value available populate with the next available value. Made use of ffill and bfill functions in PYTHON.

We couldn’t drop these NaN’s since more than 50% was NaN, also the variation in data was huge hence couldn’t use Average/Median. Previous/Next value was a safest approach.

Basic data clean up is now complete, next step is to filter and create a dataset for Ireland, since we need to analyse the Ireland livestock data.

We see that for Ireland, organic Rabbit counts were never reported:

*Pic: Rabbits for Ireland was never reported hence removing them from our analysis:*



Hence, we will drop them from our analysis.

Since we have the base data ready, all we have to do next is create a data frame that will suit our use case in inferential statistics and hypothetical testing.

While doing the ML model, we noticed that the correlation between the variable was not correct:

*Pic: Heatmap showing incorrect co-relation*

Graphical user interface, application, Teams

Description automatically generated

Hence, re-pivoted the data year wise, which showed greater co-relation and further enhanced the accuracy of the models:

*Pic: Heatmap showing improved co-relation*

Table

Description automatically generated

Sentimental analysis is done based on whether the livestock count is more than the mean value, there is a separate field created, lambda function is used, and the value evaluated.

* 1. **Statistics:**
     1. **Descriptive Statistics:**

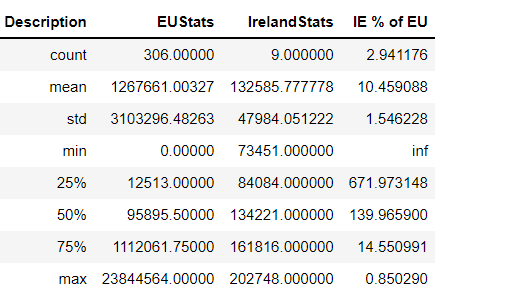
As a first step we will plot and identify the livestock which shows the highest numbers for Ireland.

Chart, line chart

Description automatically generated

The above pic shows the livestock counts over the years for Ireland, we can see that Live Poultry shows high numbers across all the years, followed by Laying Hens.

*Pic: Shows how the Ireland Statistical description varies with the EU*



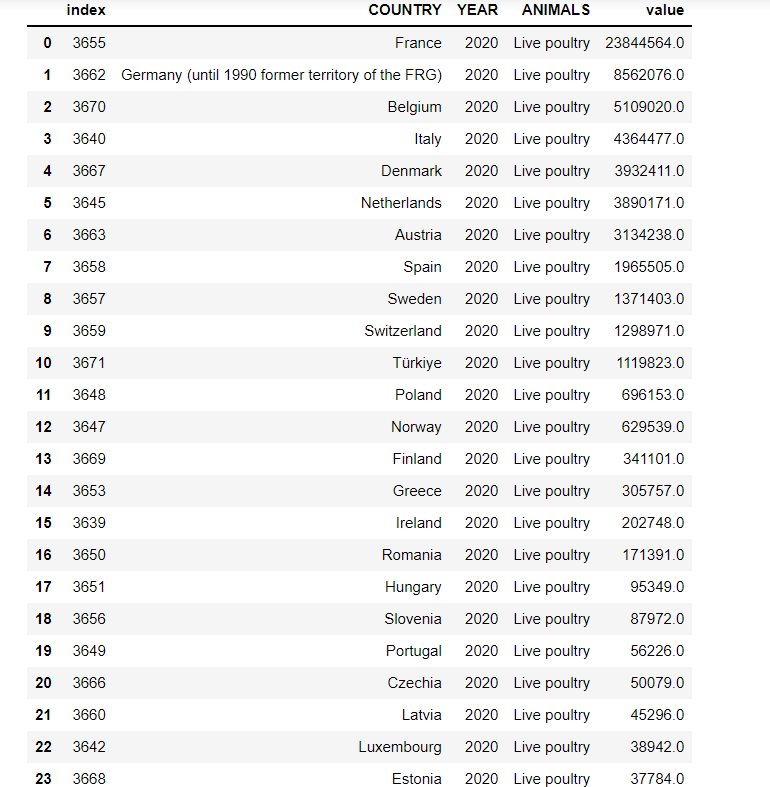
Let us focus on the mean, what it means is that of the several EU countries Ireland contributes to on an average of 10% of Laying hen counts. *(Detailed explanation along with the box plot below)*

Calculation used for the aggregate:

Text

Description automatically generated

Looking at the data for 2020, to get an idea of nearest neighbours for Ireland in terms of the laying hens count:





As highlighted above, Ireland count for 2020 is in 15th position. Now selecting the range between 70,000 and 600,000 on an average:



*Pic: Boxplot for Multiple EU Countries*

Chart, box and whisker chart

Description automatically generated

As we see in the above plot, Ireland – Romania – Hungary is comparable, and their mean values fall within 50k difference.

We see that the plots are all uneven in size, this means that the values are varying in range, the whiskers, the quartiles are all different for each country. This could be due to the topography, policy changes within the country etc.

Now, plotting the box plot for Ireland:

Chart

Description automatically generated

The box plot for Ireland shows that the range of values across the years between 2012-2020, fall within a range of 85 k to 160k,

The upper quartile group is larger than the lower quartile group, the mean is towards the upper quartile, this means that the mean is influenced by the values which are in the final 25% of the value.

**Inter-quartile range**  
The middle “box” represents the middle 50% of counts for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range. The middle 50% of scores fall within the inter-quartile range.

**Upper quartile**  
Seventy-five percent of the counts fall below the upper quartile.

**Lower quartile**  
Twenty-five percent of counts fall below the lower quartile.

**Whiskers**  
The upper and lower whiskers represent counts outside the middle 50%. Whiskers often (but not always) stretch over a wider range of counts than the middle quartile groups.

### 

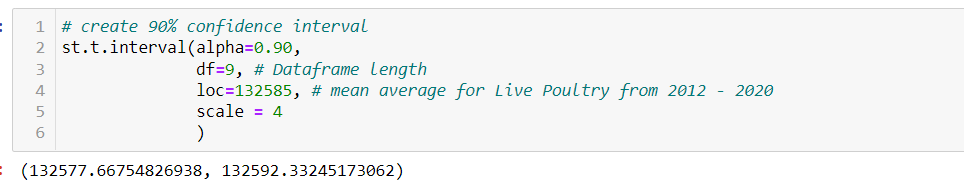
* + 1. **Inferential Statistics:**

Confidence Interval:

We know from the above for Live poultry the mean is 132585, let us now identify the Confidence Interval calculated over a range of 9 different years.

We will use t test from SciPy function from stats library.

**Case 1: 90% confidence interval:**

****

The above calculation shows for 90% Confidence Interval, the range of values expected are between 132576.94 and 132593.06 respectively.

**Case 2: 95% confidence interval:**

**Graphical user interface, text

Description automatically generated**

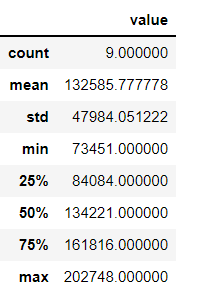
The above calculation shows for 95% Confidence Interval, the range of values expected are between 132575.94 and 132594.05 respectively.

**Case 3: 99% confidence interval:**

**Graphical user interface, text, application

Description automatically generated**

The above calculation shows for 95% Confidence Interval, the range of values expected are between 132572.00 and 132597.99 respectively.



The above cases shows that there is a wider range with 99% confidence, and we can see that over 50% is between 132572.00 and 132597.99.

* + 1. **Hypothesis Testing:**

1. **T-Test One Population**

We see above for the data(sample) from 2012-2020, the average Live Poultry count of Ireland is 132585. Also, we see from the above dataset the counts are increasing year after year, testing if the average value for the population is 200000.

#H0 : u = 200000

#H1 : u > 200000

Graphical user interface, text, application

Description automatically generated

At 95% Confidence Interval, Hypothesis is rejected because the p value is less than 0.05.

Checking if the value is greater than 200000

Graphical user interface, text, application, chat or text message

Description automatically generated

Based on the above test we can say that the average mean for Live Poultry population is over 200000.

1. **T-Test Two Populations:**

Case 1: Compare Live Poultry with Broilers, idea is to check if the live poultry is equal to Broiler count.

#H0 : u : Live Poultry Count = Broiler Count

#H1 : u : Live Poultry Count <> Broiler Count

Text

Description automatically generated

Based on the above test, we reject the hypothesis, this means the Live Poultry counts and the Broiler counts are not similar.

1. **ANOVA Test:**

Let us compare the mean variance for Live Poultry using the sample data from 2012-2019, between Ireland-Romania-Greece-Hungary.

Independent Variables created for 4 populations:

Text

Description automatically generated

Normality check:

*Probability Plot:*

*Pic: Probability Plot to check Normality*

Chart, scatter chart

Description automatically generated

The above plot shows that the data for the four countries are almost normally distributed.

Additional tests done (Shapiro Test):

Graphical user interface, text, application

Description automatically generated

H0: The variables are all normal

At 96% confidence Interval, we accept the variables are normal for the four countries.

Homogeneity of variance - Levene's test

Graphical user interface, text, application, chat or text message

Description automatically generated

The above results proves that the variables are Homogenous, this means we can now proceed with ANOVA.

**One-Way ANOVA:**

**Text

Description automatically generated**

The F value is 31.42 and p value is much less than 0.05(at 95% CI), hence we reject the hypothesis. This means that the variances of the mean value for the four countries (Ireland/Greece/Romania/Hungary) are different.

**Two-way ANOVA:**

Text

Description automatically generated

The F value is 58.59 and p value is much less than 0.05(at 95% CI), hence we reject the hypothesis. This means that the variances of the mean value for the four countries and specific years (Ireland/Greece/Romania/Hungary) are different.

1. **Kruskal Wallis Test:**

This is a non-parametric test, there is no need to check for normality and homogeneity of the variables here:

Graphical user interface, text, application

Description automatically generated

The above test shows that the pvalue is less than 0.05. Therefore at 95% CI we reject the hypothesis. Hence proving the median values of the four countries are different.

1. **U-Mann Whitman test:**

This is another non-parametric test for comparing median between two populations:

**Text

Description automatically generated**

The p-value is 0.03(<0.05), at 95% CI we reject the hypothesis that the Median values for Ireland and Romania are the same, which means the median value are different.

Taking another example, test if the median values for Live Poultry and Broilers for Ireland are the same.

Text

Description automatically generated

The p-value is 0.003(<0.05), at 95% CI we reject the hypothesis that the Median values for Live Poultry and Broilers for Ireland are the same, which means the median value are different.

* + 1. **Outcome and Challenges:**
* Identified the animal that showed steady growth in terms of its counts as Live Poultry.
* Descriptive statistics done for Ireland.
* Live Poultry included Broilers and laying hens, but t-test showed that the counts for Broilers and Laying hens are not the same.
* Several Hypothesis Testing models applied to compare between populations, outcomes are mentioned in specified sections.

**Challenges:**

* + - The data available needed lots of clean up to suit the analytical requirements.
    - Every test demanded different granularity of the data.
    - Interpretation of the data had to very detailed, since every test is for a different scenario, example: ANOVA for variances, Kruskal Wallis for Median.
    - As a beginner in statistics, had to do detailed research on the different tests, and the interpretation of the results
  1. **Machine Learning:**

The goal of this project is to predict the livestock counts using appropriate Machine Learning model.

Iterative approach was applied until the best fit model with greater accuracy was obtained. Iteration included changing the data granularity, additional data preparation, details of this is provided in the data preparation section above.

Started with a data set for every country, year, animals.

A picture containing graphical user interface

Description automatically generated

Added new variables, shifted the values to include additional variables, none of them could give better accuracy, the correlation between variables were very less.

(Details are all provided in the Jupiter notebook).

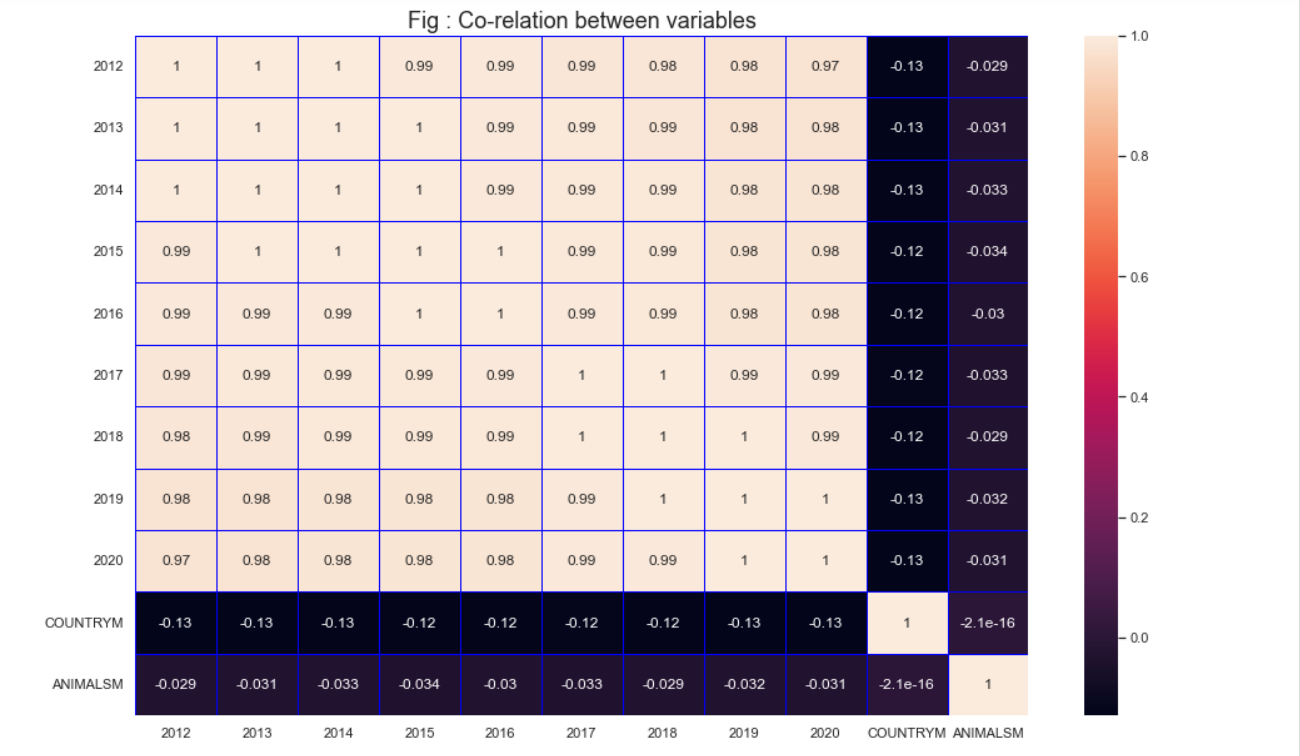
This is the point, where I decided to change the granularity of the dataset:

Graphical user interface

Description automatically generated with medium confidence

*(Made use of PYTHON pivot function to achieve this)*

Year Variables showed higher co-relation:



I had two approaches to the problem, use 2012-2020 as all independent variables and apply unsupervised learning, or use 2012-2019 as independent and 2020 is the target for prediction. I chose the second option.

X = Values for Years from 2012-2019, also included COUNTRTM, YEARSM (these are mapped integers corresponding to the country and year).

Y = Target Variable to predict 2020.

**Test Train Data Split:** We will split the test and train data with 30% of the total data in test. We will use the same set of data across all the models and identify the model showing the best accuracy/r2 score.

* + 1. **Apply Regression Models**

***Linear Regression:***

The first model that was applied was the linear regression model.

The regression models are applied to the continuous dependent variable. Linear regression is applied when the independent and dependent variables are linear.

The formula used is

**ŷ = w0 + w1x**

The goal is to find the best value for **w0** and **w1.**

First step we will divide the train and test data:

Chart, scatter chart

Description automatically generated

Apply Linear regression, below is the outcome:



Chart, histogram

Description automatically generated

Above graph shows train and test are very close.

*Pic: Prediction based on test and train data*

Text

Description automatically generated with medium confidence

**K-Fold Cross Validation:**

**Graphical user interface, text, application

Description automatically generated**

K-Fold CV showed accuracy of 0.968 and 0.933.

**GRID Cross Validation:**

Applying GRID CV, to check the optimal value of features:

**Graphical user interface, application

Description automatically generated**

*Pic: Table showing rank of features*

**Table

Description automatically generated**



As seen above, best rank (1) is when the number of features is 3.

**K Neighbour Regressor:**

Number of Neighbours considered=10



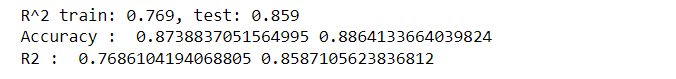
Train accuracy and Test accuracy are at 87.3% and 88.6%

This is less than what we saw for linear regression with GRID CV.

**Random Forest Regressor:**

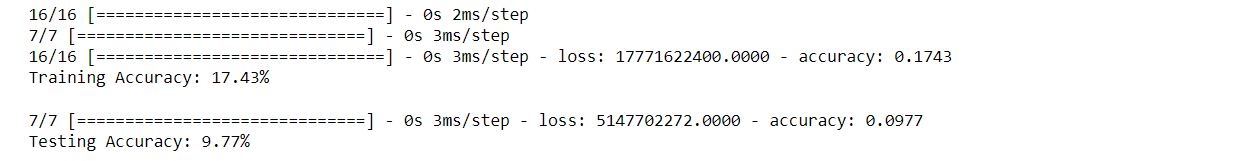
This makes use of the RandomForestRegressor object within the ensemble class of the sklearn library.

The X variables are the five variables we saw in the box plot above showing high co-relations, y is the dependent variable.

****

**Artificial Neural Network**

Created a Neural network with 4 hidden layers to get optimal accuracy:



Accuracy seen is very less.

Chart, line chart

Description automatically generated

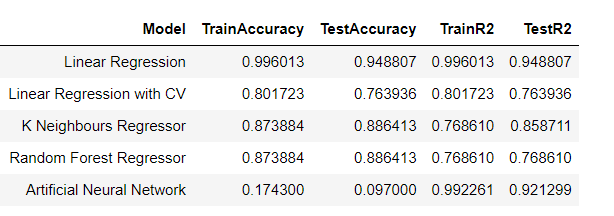
Training accuracy is optimal for Epoch = 1 and reduces further and then increases at 8 and then further at 10.

Test accuracy is lesser and consistent.

The overall accuracy is very less for ANN.

**SUMMARY**

Below table summarizes the Accuracy and R2 scores of different models used.



Linear Regression is a clear choice based on the analysis done here with a accuract of 99% (Train) and 94%(Test), followed by Random Forest .

ANN shows very less accuracy.

***Pic: Comparison of ML Models***

Graphical user interface, application

Description automatically generated

f

* + 1. **Sentimental Analysis:**

Sentimental Analysis are done here using the existing numerical data. This means there is no need for additional Feature Engineering techniques like the Count Vectorizer.

Aim of the analysis is to analyse the sentiments based on the live counts of each animal for each year and country in comparison with the average count for each year and country.

Multinomial library is used for the prediction:

*Pic: Classification Report*

Table

Description automatically generated

Classification Report shows a negative sentiment probability is with 96% precision, positive is with 76%.

**Logistic Regression:**

Cross Validation score: 0.86

GRID CV score: 0.86

We can conclude that sentimental analysis with the above classification report is 86% accurate.

* 1. **Visualisation:**

Usage of appropriate charts and tables are implemented across the project. This includes

1. **Acknowledgments**:

The idea for the structure and details to be included in the project came from the research paper “Traffic Volumes Prediction using Big data Analytics “.

1. **References:**

Template supplied by the CCT College.

Class notes shared by the lecturers for getting the PYTHON code, visualization, ML codes, and Statistics.