Database and Analytics Programming Project

Bryan O'Donohoe[#1], Ermesa Pepe[#2], Marcin Starzyk[#3]

[#]*School of Computing, National College of Ireland*
Mayor Street Lower, International Financial Services Centre, Dublin, Ireland
[1]x20212828@student.ncirl.ie, [2]x20212887@student.ncirl.ie,
[3]x20212836@student.ncirl.ie

***Abstract*— This paper demonstrates usage of the common programmatic tools and technologies (Python, Jupyter), as well as cloud based infrastructure (MongoDB, PostgreSQL, GitHub) for data processing. .**

Using publicly available datasets as an example this document presents the full process of data analytics following KDD methodologies : selection of dataset; programmatic retrieval of data; data cleaning and preprocessing; data reduction; analysis and visualisation of data.

***Keywords*— data analytics, FIFA; Meat Consumption; Health; Covid; BMI;**

I. Introduction

Data analytics is known as the set of methodological principles and multidisciplinary techniques to interpret and extract knowledge from data.

Data Analytics methods are based on techniques derived from various disciplines, mainly mathematics, statistics, computer science and social sciences, particularly in the subdomains of databases and data visualisation, artificial intelligence or machine learning.

In detail, five different phases of data analytics can be identified as shown below in Fig. 1, although, in the following project, we will mainly focus on the first three phases. These phases should be completed in succession, however the process can become iterative as discoveries are made at a later stage.

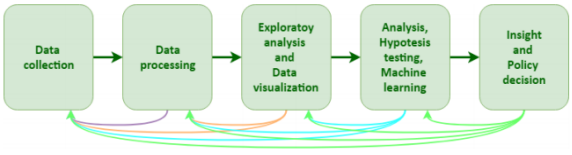

Fig. 1 Flowchart of Operations

A. Project Objectives

Projects Objectives can be summarized in the following main points:

1. Identification of suitable datasets and programmatic retrieval
2. Programmatic storage of data in appropriate database prior to processing
3. Programmatic pre-processing, analysis and visualisation of data
4. Programmatic storage of the processed output data in the appropriate databases
5. Programmatic creation of dataset by joining together datasets

Project objectives are to be achieved as a team effort from the workflow perspective objectives can be divided into 2 stages:

Stage 1. An Initial process and analysis of each dataset independently of the other datasets. (points - 1-3)

Stage 2. The datasets are programmatically merged, and further analysis is conducted on the merged datasets. (points - 4-6)

B. Motivation

Motivation for the project is to exhibit (and broaden) the programming skills and knowledge gained in the Database and Analytics Programming Module.

We will show how the project's scope can be successfully achieved using a powerful and general-purpose coding language like Python, which is largely used in data science and writing system scripts. Among other things, Python provides an extensive range of supporting libraries like NumPy, Pandas, Matplotlib, Scikit-Learn, Scipy.

Besides the programming languages, the utilisation of databases like MongoDB and PostgreSQL will facilitate the data preparation and exploratory statistical analysis for the datasets.

C. Research Question

Research question selection (which is bound to selected datasets) is de facto a pretext to fulfill the project objectives.

The topic chosen for the scope of analysis could help us to formulate the various problem statement and answer to questions similar to:

- What is the risk of COVID-19 infection from animals or animal products imported from affected areas?
- How will the COVID-19 Pandemic shape the future of meat consumption?
- Should our health make us rethink our relationship with meat?
- Is the coronavirus pandemic in general related to meat production and consumption and lifestyle?
- Can overall player rating be predicted in FIFA 21 using other attributes?

II. Related Work

A. Literature Review

At the moment, there is no scientific evidence that any animals or animal products authorised for entry into the European Union pose a risk to the health of citizens as a result of the presence of COVID-19.

Based on past academic research on the most recent zoonotic viruses (SARS-CoV-1 virus, H5N1 bird flu, swine flu) it is reasonable to spend time analysing if it exists a relation/correlation of meat production/consumption to our

health, considering it as a vulnerable factor beside other factors like obesity or sedentary lifestyle.

It has been found that Coronavirus indicates transmission risk increases along wildlife supply chains for human consumption. Considering that eating meat is proven to be linked to a wide range of illnesses. Zoonotic viruses are the highest correlated to the mass global production and distribution, started by the high demand and consumption.[1]

The topic is also affecting public opinion. Notably, the trends during the COVID-19 Pandemic show a decrease in meat and seafood consumption.

B. Previous Analysis

Still, research has to be conducted on the evidence that meat can be a vector to spread the covid-19. Until now, the available data is very limited, and the analysis conducted is mostly based on forecast data using projections based on the Aglink-Cosimo model.

Aglink-Cosimo is a recursive-dynamic partial equilibrium model developed and maintained by the OECD and FAO to formulate deterministic projections for the analysis of specific market uncertainties [2][3].

In the following report, we will also illustrate the support of the Python libraries available and a comparison between the ARIMA forecast model and the *Aglink-Cosimo* model already available on the OECD-library website [4]. We can compare the two forecast models trends using Python to show the pandemic's effect on meat consumption and production in the next ten years.

Different research has been conducted around the vulnerability assessment for pandemics:

The conceptual framework of Urban Vulnerability Assessment (UVA) for Pandemics is illustrated in the figure. The assessment involves four main stages. The first stage is the identification of vulnerable factors influencing Pandemics, see. The second stage is to transform the raw input data from each vulnerable factor into a probability distribution. The third state groups geographic areas with similar characteristics into classes to assign a vulnerability level. After that, an aggregation method is applied to create a unique rank for each class see, where a higher rank is assigned to a higher vulnerability level.

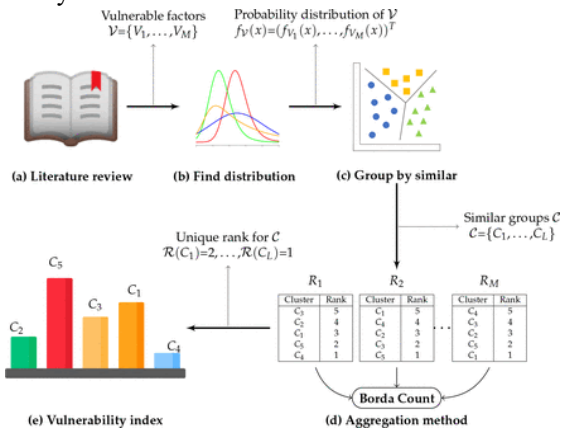

Fig. 2. Flowchart of operations [5]

Our research could be further developed and be included in the lists of studies already developed in the identification of more vulnerable factors for Pandemics. The conceptual framework of Vulnerability Assessment for Pandemics as outlined in Fig. 2 above.[5]

Several types of research have been conducted and studies that investigate the possible vulnerable factors related to Pandemics. The studies consider both factors related to past Pandemics (i.e., 1881 Fifth cholera, 1918 Spanish flu influenza, 1957 Asian flu influenza, 2003 SARS, 2009 h1n1, 2013 West Africa Ebola) and factors found in the current COVID-19 Pandemic. The search retrieved studies for which the study's title, abstract, or keywords indicated the study examined a type of vulnerability in Pandemics. Then, a manual assessment is made for every study against eligibility criteria and statistical analysis is conducted to create a vulnerability index. [5]

Analysis of the COVID-19 pandemic previously reviewed the global economic impact by various factors relating to market sectors [6]

Media and technology use predicts ill-being among children, preteens, and teenagers independent of the negative health impacts of exercise and eating habits.[7]

Analysis of FIFA 19 Playstation game was previously carried out [8] in which attributes were analysed for highest ranking players and relationship between overall rating and specific attributes. However, no analysis was carried out to indicate the spread of the key attributes across the globe.

III. Methodology

FIFA Dataset[#1]

A. Description of the underlying dataset(s) and justification of choosing them.

Player information relating to each avatar in the FIFA 21 Playstation game was chosen as the data for this analysis. All records for the 18,000 player avatars were extracted from the website sofifa.com. This method of data collection and source of data was chosen due to a similar method used for previous versions of the game on this open source data [8]

The data records for each player listed attributes for their avatar such as overall score, height, weight, skill abilities, nationality and many more such that we have 93 columns for each record.

B. Descriptions and justifications of the data gathering and handling activities carried out (e.g., use of APIs, databases, etc.)

First, a request URL provided scrapes some basic player information off the main webpage. The "ID" field collected is then used in conjunction with a loop to iterate through each player's page and obtain more detailed information. This data was sent to a database to store before cleaning.

C. Descriptions and justifications of the implemented data processing algorithms

In order to combine with the other datasets in the project, nationality was chosen as the key variable when creating a relational database.

An ETL (extract, transform and load) tool (cleaning.py) was scripted in Python for the data processing. Using libraries such as pandas and numpy which are built specifically for handling datasets such as the one in question here, the tool was developed to clean and format the data such that it could be used in conjunction with the other datasets in the project. There is a limitation to this method however as the tool stores the dataset in memory during processing which would make the solution hard to scale. For the purposes of this analysis it was sufficient.

Fig. 3 gives an example of the type of cleaning that was required for some of the variables.

```
# Release clause currently in the format of "€1.2M" or "€300K"
# Two functions created to transform this into numeric format

def release_clause1 (value):
    try:
        return value.split("€")[1]
    except:
        return value

def release_clause2 (value):
    try:
        return float(value.split("M")[0])*1000
    except:
        return float(value.split("K")[0])

df["release_clause_clean"] = df["Release Clause"].apply(lambda x: release_clause1(x))
df["release_clause_clean"] = df["release_clause_clean"].apply(lambda x: release_clause2(x))
```

Fig. 3. Function created to transform Release Clause into numeric format

Data processing included NAN checks and fixes, lambda functions for conversion of strings to numeric format and lambda functions for bucketing of countries.

D. Justifications for the choice of technologies used (i.e., programming languages, databases, etc.)

Python (through Jupyter Notebook and Google Colab) was the programming language of choice for the processing, storing and visualisation of all data. As previously mentioned, the packages available as well as the wealth of resources online, meant that the language would yield optimum results.

Cloud based architecture was selected to enable collaborative work of geographically dispersed team members.

MongoDB was used to programmatically store the data prior to processing. Github was used for version control. Postgres was used for storing the processed datasets such that they could be accessed for analyses.

Meat Dataset[#2]

A. Description of the underlying dataset(s) and justification of choosing them.

In the last 40 years, meat production and consumption have increased considerably and significantly impacted human health and the whole environment. It is not new that meat, as every processed food, requires several resources to be produced, such as water, energy, feed, and that produces an amount of emission of CO2 that is directly proportional.

Besides meat consumption, pandemics' vulnerable factors could also be related to meat production and meat demand, defining a real *"meat issue"*, as Kendra Pierre-Louis, an expert climate change journalist, described it in the New York Times [9] Therefore, it is not just about greenhouse gas emissions, but the entire food production cycle. The latest data analysis shows that livestock contributes about 5% of the nearly 37 gigatons of carbon dioxide emitted by human activity into the atmosphere each year.

Based on current rates, 100 years of carbon dioxide production through cattle processing would have a minimal impact, leading to an increase in temperatures of around 0.1 degrees. Furthermore, meat production adds about 0.15 gigatons of methane and 0.0065 gigatons of nitrous oxide to the atmosphere each year. [10]

B. Descriptions and justifications of the data gathering and handling activities carried out (e.g., use of APIs, databases, etc.)

The following section will illustrate how the data have been gathered and transformed for the project's scope. The project folder available on GitHub (https://github.com/BryanOD95/DBAP/tree/main/ermes) mainly contains files showing all the data Extract, Transform, Load(file enumerated from 0 to 6).

The original Dataset available at http://www.fao.org/faostat/en/#data/FBS consists of two separate datasets containing many commodities and commodity aggregates, among which only the ones related to meat commodity are selected as outlined in Fig. 4.

Data	Food Balance Dataset	
	Unit Measure	*Time Coverage*
Food Balance1	Export Quantity[1000T] Import Quantity[1000T] ProductionQuantity[1000T] Food Supply Quantity[kg/capita/yr]	**2014-2018** Size 7.2 MB Rows:292.839 Columns:12
Food Balance2	Export Quantity[1000T] Import Quantity[1000T] ProductionQuantity[1000T] Food Supply Quantity[kg/capita/yr]	**1961-2018** Size 25.49 MB Rows:238.560 Columns:60

Fig. 4. Dataset Description

Data collection is the set of techniques and methodologies for extracting useful information from large amounts of data. The following sections describe in more detail the operations performed to clean the data in such a way as to make it usable in subsequent operations. Overall the data has been standardised, validated, and scrubbed before storing it in the final database *"dap"* in PostgreSQL on the Cloud.

C. Descriptions and justifications of the implemented data processing algorithms

The original datasets contain many columns and categories; for the project, the five available meat categories contained in the column *Item* are selected: *Bovine meat (including buffalo and beef), Pig Meat, Poultry Meat (including chicken, turkey), Mount and Goat Meat, Other meat (Mule Meat, Camel Meat, Horse Meat)*. And four typologies of service from the column *Element*: *Meat import, Meat export, Meat production, Meat consumption*.

The programming language chosen for the scope of extracting, processing, analysing, and displaying data is *Python* in version 3.9.1; the choice of a dynamic and flexible language such as Python has made it possible to simplify very complex operations that otherwise would have required more time and more code. In this case, Python is used to programmatically retrieve a semi-structured dataset CSV, using the function *read_csv()* and storing the data in MongoDB in *JSON* format using lambda expression and *insert_many()* function. Python includes different libraries that have been used for the project: *NumPy, SciPy, Pandas, JupyterDash, Matplot, Seaborn*. [11]

D. Justifications for the choice of technologies used (i.e., programming languages, databases, etc.)

The datasets are stored in *MongoDB Atlas* that offer the advantage of a flexible, scalable cloud service accessible anytime and anywhere.

Reading the data. The best solution for reading the data has been using *Pandas* that offers many commands for reading data regardless of the file structure that contains the data (*CSV*, *Xls*, *SQL*). Generally, the data is read and inserted in a *Pandas DataFrame*, which can be modelled according to the needs of subsequent operations. With Pandas, it is possible to rename rows and columns and select a specific column that will constitute the row index and compute statistical calculations. In the case of missing data, in Pandas, it is possible to identify them. Pandas, in its functionality, can avoid, replace, or delete rows and /or columns that contain null parameters identified with the *NumPy* value *NaN*.

Pre-processing, transformation. The pre-processing phase aims to obtain data that present the **five characteristics of quality data: Accuracy, Validity, Completeness, Consistency and Uniformity**.

Data quality is a fundamental aspect and can be compromised by: *Missing values, Noise(data distortion), Outliers, Duplicates*. [12] Data wrangling operations begin with analysing the project requirements from which valuable data can be identified, **excluding data not included in the scope of the analysis**. The main problems identified and solved during the project are missing data, unnecessary parameters, replaced appropriate values when possible, removed duplicate values, removed irrelevant/unnecessary observations, fix structural errors, fixed irregular format for dates and names, checked for the presence of unwanted outliers, handle missing data and NaN values. As shown in the file *1.data_clean_azure.ypynb,* the stages have been iterated several times.

Fig. 5. Data Exploratoration [13]

The file *1.data_clean_azure.ipynb* shows the entire iterative process, including the first filter stage, mainly consisting of filtering:

- Five meat categories: Bovine Meat, Meat Other, Mutton & Goat Meat, Pigmeat, Poultry Meat
- Four services: Meat Production, Meat Consumption, Meat Import, Meat Export filtered among the list in the element column in both datasets

TABLE I.
DataFrame size before/after filtering

Data	DataFrame Size Before/after filtering	
	Before	*After*
df1.shape	(292839,12)	(3604,12)
df2.shape	(238560,60)	(3644,60)

Missing data. The two datasets have different numbers of rows due to discrepancies in the number of countries and missing data related to some meat services. Due to the extensive quantity of data it can be beneficial to create a graphical representation of the missing data, using a *heatmap* and the library *Seaborn*. Seaborn is a Python data visualisation library based on *Matplotlib* very useful to build statistical graphs:

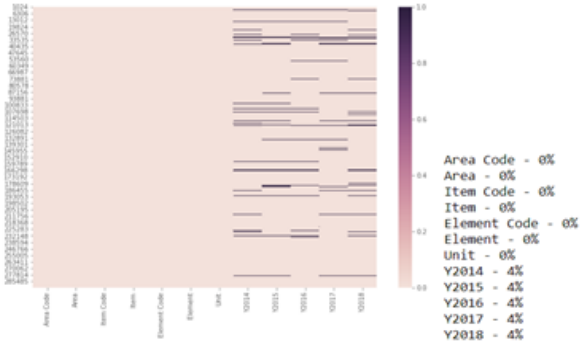

Fig. 6. Missing data

Transformation. The column *Year* has been modified, eliminating the Y letter from the name (i.e. Y1961 to 1961) to be ready for the successive reshaping.

Further filtering for uniformity. For uniformity and consistency it is decided to exclude countries that do not show data for all the five meat typologies and four meat services. Once the filtering process has completed, the result is two DataFrames respectively: *df1*(2544 x 12) and *df2* (2544x65), containing the same list of 106 world areas, that are merged in a unique frame named *final*(2544x65), using the function *merge()*.

Handling a time series dataset, the data grouping function is beneficial. The possibility of grouping data allows to carry out operations such as sums, subtractions, counts, and various statistical calculations whose results are entered in specific columns with the operation's name; for example, the column *All* contains the sum of all the meat categories. Among the various operations that can be performed on the data, tere is the so-called **pivot tables** , which are tables with an index, the expansion of the column elements as columns, and the linked values to other columns as table values. This kind of table is handy for the graphical representations that will then be carried out. The pivot-tables are inserted into the Dataframe dfP(24593 × 6).

Year	Area	Area Code	Item / Element	Bovine Meat	Meat, Other	Mutton & Goat Meat	Pigmeat	Poultry Meat	All
1961	Africa	5100	Consumption	6.91	1.85	2.69	0.68	1.31	13.44
			Export Quantity	89.00	10.00	1.00	5.00	1.00	106.00
			Import Quantity	64.00	4.00	12.00	26.00	9.00	115.00
			Production	1898.00	508.00	717.00	163.00	346.00	3632.00
	Algeria	4	Consumption	4.06	0.49	3.63	1.41	1.81	11.40
...	...	...	...	...	...	...	...	...	...
2018	Zimbabwe	181	Consumption	8.50	2.57	1.27	1.25	4.83	18.42
			Export Quantity	0.00	0.00	0.00	0.00	4.00	4.00
			Import Quantity	1.00	0.00	0.00	0.00	4.00	5.00
			Production	111.00	37.00	26.00	18.00	66.00	258.00
All				15295834.28	1303531.75	2773577.93	20028864.35	14560909.34	53962717.65

24593 rows × 6 columns

Fig. 7. Pivot Table

From the pivot table (Fig. 7), all the data related to the continents have been filtered and stored in a different data frame dfC. The final data is stored using the function insert_many() in a PostgreSQL Db installed on an Azure VM with Debian OS, accessible from the public through SSH. A local VM with Debian OS has also been used, enabling queries to be performed faster than on the remote VM in the Cloud.

COVID and BMI datasets[#3]

A. Description of the underlying dataset(s) and justification of choosing them.

Description of each of the three datasets selected are presented in Tables II-IV

TABLE II.
COVID DATA PER COUNTRY

Dataset	Covid Data
Description	Basic Information about Covid information per country (*confirmed, deaths, recovered, active*)
url	https://covid2019-api.herokuapp.com
Access Method	API call
Format	JSON
Dimensions	192x5

TABLE III.
POPULATION DATA PER COUNTRY

Dataset	Countries Population
Description	Contains basic information about Countries - among other (*Population*, *Land Area*, *Density*)
url	https://www.worldometers.info/world-population/population-by-country/
Access Method	Web Scraping
Format	lxml
Dimensions	235x11

TABLE IV.
OBESITY DATA PER COUNTRY

Dataset	Obesity Data
Description	Contains information about prevalence of obesity among adults (presented as % of population with BMI>30), provides breakdown of obesity index for male, female, and both sexes for years 1975-2016.
url	https://apps.who.int/gho/data/view.main.BMI30Cv?lang=en
Access Method	HTTP GET of xml file
Format	xml
Dimensions	195x126

Datasets were purposefully selected to allow for realisation of projects objectives. Despite low dimensions they do allow for demonstration of a wide range range of data processing techniques. Datasets' Format and Access Methods are complementary to datasets selected by other team members as shown in Figure 7.

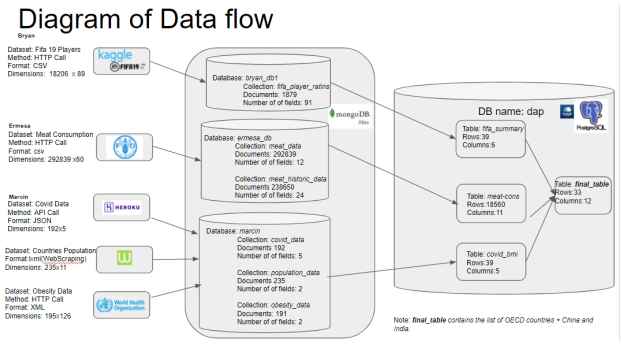

Fig. 7. Diagram presenting follow of data (full size version in Project Presentation file)

B. Descriptions and justifications of the data gathering and handling activities carried out)

Gathering data and initial part of processing of the datasets can be found in Jupyter notebook file: Retrieving and Storing Data in Atlas MongoDB.ipynb. File contains main function which retrieves data from all 3 sources and stores them in MongoDB

Retrieval, initial processing, and storing in MongoDB

- Covid_data: Covid data is retrieved via API calls which is critical since Covid information is changing on a daily basis. Since data retrieved is in JSON format, they are stored in Atlas MongoDB instance without any processing.
- Population_data: Retrieval from wordometer website is done using WebScraping. Since the Wordometer table contains 11 columns, initial processing is to select only columns 'Country or Dependency' and Population (2020). Before storing in Atlas MongoDB- names of columns are changed to 'Country' and 'Population' respectively.
- Obesity_data: dataset from WHO website is downloaded and stored locally in temporary file response.xml. From the file - out of 126 columns 2 are selected. (Country and Obesity_percentage for 2016. Rows with NaN values are discarded . Data is stored in Atlas MongoDB

Storage in the cloud based Mongo DB instance limits processing needed (one of the files is JSON formatted), and at the same time provides constant access to data by other team members.

Retrieval from MongoDB and storage in PostgresSQL

The next part of data processing Jupyter notebook file: *Retrieving Data from Mongo DB- processing and storing in PostgresSQL.ipynb.* For the list of OECD+ countries all fields are retrieved from *obesity_data* and *population_data* collections - while from *covid_data* only the column with the number of deaths is downloaded.

Once all data are in one data frame an additional column *covid_death_precentage* is added and data frame is stored locally in csv file to allow for local statistical analysis (see file *Sample of statistical analysis in R.pdf)*

As a last step dataframe is uploaded to cloud hosted PostgresSQL DB instance for aggregated processing of datasets from all members.

C. Justifications for the choice of technologies used (i.e., programming languages, databases, etc.)

Cloud based architecture was selected to enable collaboration between geographically dispersed team members (see: Fig. 8)

Technological stack

- MongoDB: hosted in *MongoDB Atlas*
- GitHub: hosted in GitHub
- PostgresSQL - hosted in *cloudclusters.io*)
- Anaconda/Jupyter
  - Python (main modules: *panda*, *pymongo*, *ElementTeree*, *sql alechemy)*
  - R

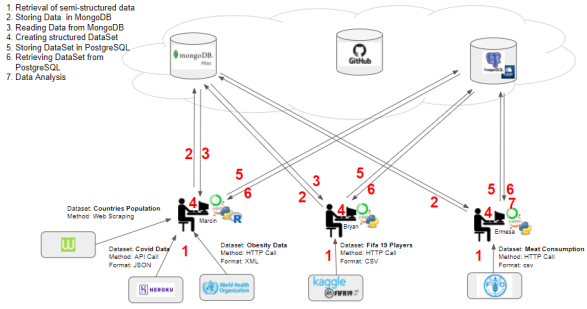

Fig. 8. Diagram presenting architecture of solution (full size version in Project Presentation file)

D. Presentation of results by making appropriate use of figures, tables, etc.

To demonstrate agility of environment statistical analysis/visualization have been done in R language. Results can be seen in file *Results were processed as per Sample of statistical analysis in R.pdf.* Samples of two graphs are presented in Fig. 9. and Fig. 10. No correlation between obesity and Covid rate has been noticed.

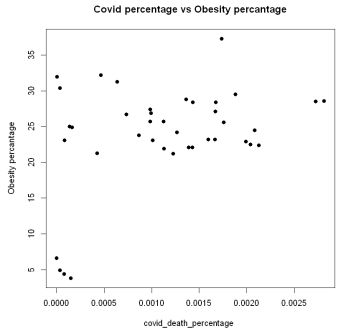

Fig. 9. Scatter plot - obesity vs Covid Rate for OECD countries

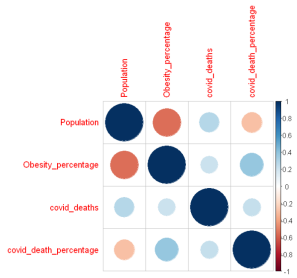

Fig. 10. Correlation between values in columns of *covid_data*

E. Evidence of how the project objectives were met

Evidence of meeting projects objective can be found in online repositories and DBs:

- Github: code and
- MongoDB and PostgresSQL DB

as well as the result section of this document.

IV. Results

FIFA Results[#1]

A. Presentation of results by making appropriate use of figures, tables, etc.

Fig. 11 below details the correlation between the overall player rating and the other skill attributes of the avatar.

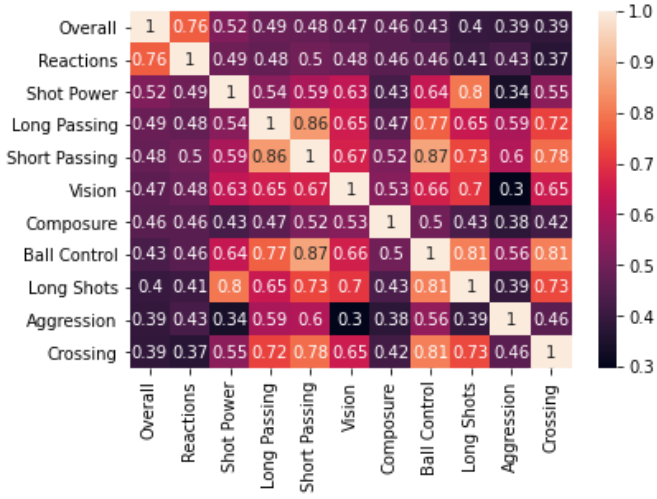

Fig. 11 Correlation Matrix of Player Attributes

Fig. 12 below displays the feature importance from a Gradient Boosted Machine for predicting the overall player rating using the other key attributes as factors in a regression model.

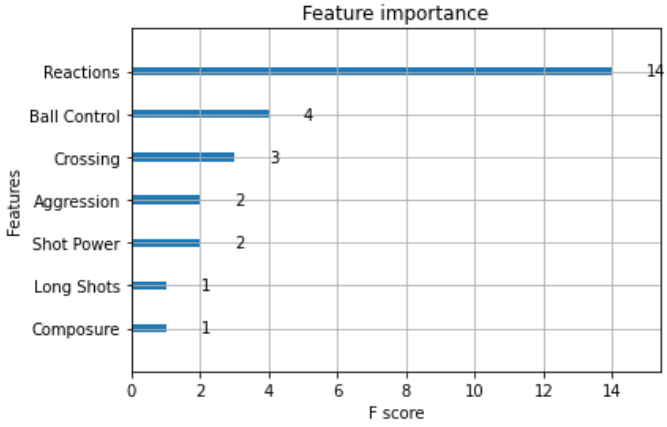

Fig. 12. Feature Importance - GBM output

Fig. 13 below details the average player reaction rating for each nationality in the FIFA 21 game.

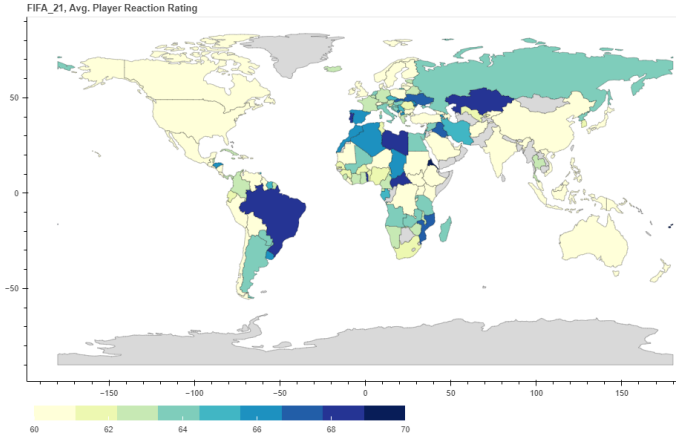

Fig. 13. World map of average player reaction rating in FIFA 21

Table V below details the player with the maximum value for a given attribute.

TABLE V
Player with maximum attributes

	Name
Overall	Lionel Messi
Reactions	C. Ronaldo dos Santos Aveiro
Shot Power	Aleksandar Kolarov
Long Passing	Kevin De Bruyne
Short Passing	Kevin De Bruyne
Vision	Lionel Messi
Composure	Lionel Messi
Ball Control	Lionel Messi
Long Shots	Lionel Messi
Aggression	Klaus Gjasula
Crossing	Kevin De Bruyne

B. Evidence of how the project objectives were met

Section 3b outlines how the Dataset was collected using web scraping from a suitable website. It was then programmatically stored in an appropriate database. ETL was

carried out prior to visualisation of data as per section 3c using a wide range of functions and transformations. Postgresql was utilized for the combination of this dataset with the rest of the group.

C. Discussion of the research findings, their interpretation(s) and implications

Fig. 27 above illustrates the correlation between the overall rating of a players ability with each of the top 10 attributes. As shown above, reactions are the highest indicator of player ability.

Fig. 28 goes one step further than the correlation matrix in ranking reaction as the most important feature in predicting overall rating. It is notable that Ball Control is ranked next highest which would not be evident from the correlation matrix in Fig. 27. It is interesting that while reactions would be considered an inherent trait, ball control would be a trait that could be improved through conscious practice.

Fig. 13 details a world map containing the average reaction rating of each country. Brazil, Portugal and Spain are not surprisingly among the highest scores in this analysis (excluding countries with numbers too small for credibility)

Table IV outlines the highest scoring players for each of the top 10 attributes. Lionel Messi is the highest rated player in the game and it is no surprise that he also scores the highest in 4 of the 10 attributes analysed.

Meat Results

A. Presentation of results by making appropriate use of figures, tables, etc.

Since much of the data is difficult to understand as they are in large quantities, and because often without being able to compare them with others, it is difficult to understand their meaning, it is very effective to take this data and use it for the graph construction. Several libraries allow the representation of data in Python, those used for this project are *Matplotlib*, *Seaborn* and *Plotly*, *JupyterDash, Altair*. The aim is to focus on using Python libraries that can give meaning to data in a few lines of code. The data analysis operation, in fact, is very complex without a graphic representation of the same. Analysis of the meat production and consumption per continent:

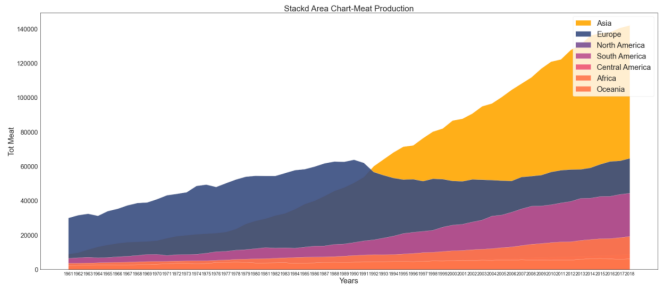

Fig. 14. Meat Consumption by World Regions

With *Matplotlib* and *BarRace* it is possible to create an animation that shows the evolution in time of meat production has impressively increased in the last 50 years.

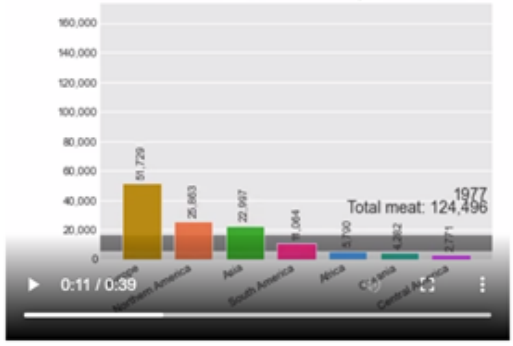

Fig. 15. Meat production in 1979

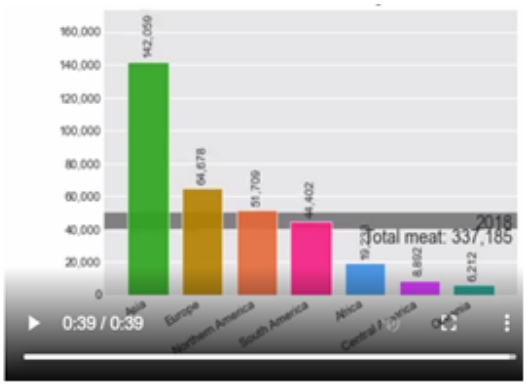

Fig. 16. Meat production in 2018

We can notice since 1977 how the *Asian* continent has moved from third place to first place in Meat Production. *Europe* and *North America* were the dominant meat producers. In 1961, Asia produced only 12%. By 2013, Europe and North America's share had fallen drastically.

Meat production by meat category(i.e. bovine and poultry):

Fig. 17. Bovine meat production

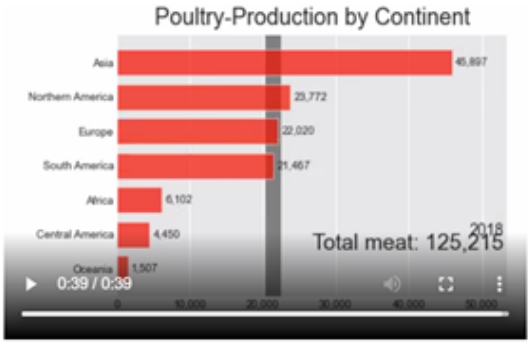

Fig. 18. Poultry Meat Production

Asia is still the top meat producer of Poultry, Bovine and Pig meat.

Meat Consumption by Continent:

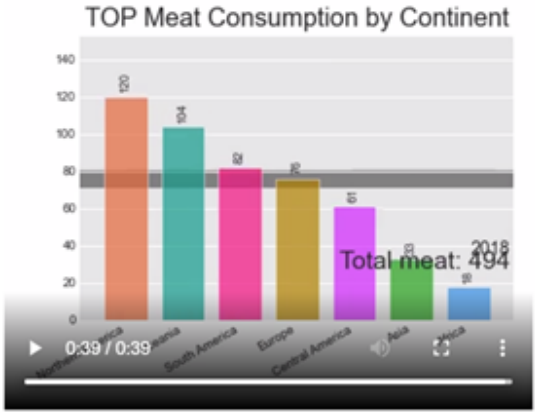

Fig. 19. Meat Consumption

North America is the highest meat consumer in 2018 followed by Oceania and South America.

To understand the correlation between the various services, it is useful to create a dashboard. Among the benefits of using dashboards: all data is grouped more simply; interaction with the data to better absorb more information faster; identify patterns and relationships in the data; outliers, peaks and other inconsistencies can be more easily identified. Below an example of one of the dashboards created using ***Jupyter-Dash*** to compare different meat services for each year 1961-2018.

The dashboard is available at "*http://127.0.0.1:8050/*" once loaded:

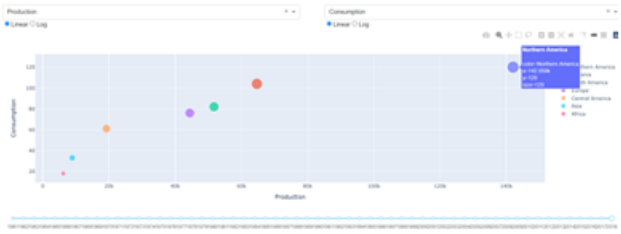

Fig. 20. Dashboard Example

Treemap is also used to display the data as it optimises data visualisation. The data is displayed as rectangles of various sizes and colours, which can contain other smaller rectangles. Treemap related to meat production and consumption in 2018:

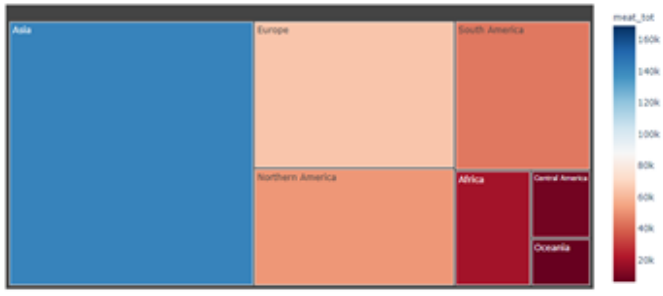

Fig. 21. Meat Production Treemap Chart

Fig. 22. Meat Consumption Treemap Chart

Asia is the largest producer of meat globally, while it is only in fifth place as a consumer, indicating that most of the meat is probably exported to demanding countries. North America is first place as a meat consumer with an average of *120(kg/capita/year*) of meat consumed. Oceania is the 2nd continent for meat consumption but the last in meat production they have a high amount of imported meat quantity.

It is interesting to analyse in more detail which are the countries that contribute most to meat production and consumption. Meat production per country:

Fig. 23. Meat production- Top countries

The United States is the world's largest bovine meat producer, producing more than 12 million Tonnes in 2018. Other major producers are Brazil and China.

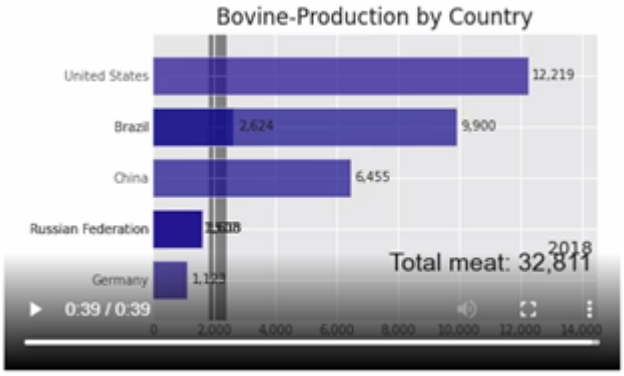

Fig 24. USA Biggest Bovine meat production

China is the world's largest, producing around 55 million tonnes of pig meat in 2018. Chinese pig meat production increased rapidly in the last 20 years, from 1.5 million tonnes in 1961 to 54 million tonnes in 2015. The other major producers include the United States, Germany and Spain and Brazil.

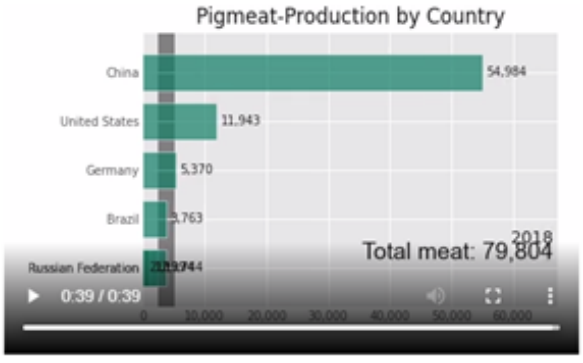

Fig. 25. China-Biggest Pigmeat producer

Dashboard to compare various meat service:

Fig. 26. Dashboard Export vs Production- China

We can notice that from 2001 to 2018 the meat consumption in China is almost the same while the production is almost double, this could mean that most of the meat produced is exported.

In China, since 1990 the country shifted from the main pig meat producer to a considerable quantity of poultry and beef due maybe to a change in wellness among the global population demanding more variety of food and white meat.

We can note that India is one of the biggest beef producers even if Indians are not consumers of such products for religious reasons.

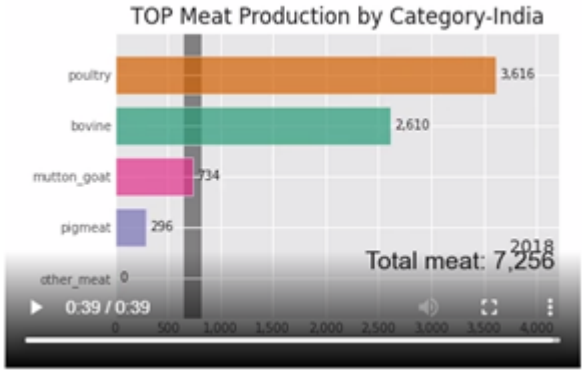

Fig. 27. Quantity Meat production by Category(1000 Tones)

This fact may suggest that Beef meat is mainly produced to be exported to countries with the highest consummation demand, such as the USA and Europe.

V. Conclusions and Future Work

A. Conclusion Bryan

As outlined previously there is the ability to fit a regression model to the overall rating of a given player using some of the skill attributes.

The model here was limited to 10 skill attributes that contained numeric data. In future work, categorical data such

as body type, position or international reputation could be used to reduce the residuals of such models.

Further analysis could be carried out on the geographic features by using K-nearest neighbours to smooth the averages and allow the fitting of a location factor in such a regression problem.

B. Conclusion Ermesa

Matplotlib, Seaborn, and Plotly are unique in their way. Also, a few other libraries like Altair, ggplot, and Pygal have been tested and are suitable for data visualisation. Choosing one among them depends on personal needs and preferences. Seaborn has been very useful for static views and Plotly, Jupyter-Dash for highly interactive web-embedded views.

The graphs and dashboards created are a picture of the historical moment we are living in. The population number is growing in many countries, such as China, and the demand for meat is increasing rapidly from more advanced countries. Still is evident a considerable difference with developing countries.

Brazil is one of the largest beef exporters globally, mainly destined for the Chinese market and Hong Kong. [14]

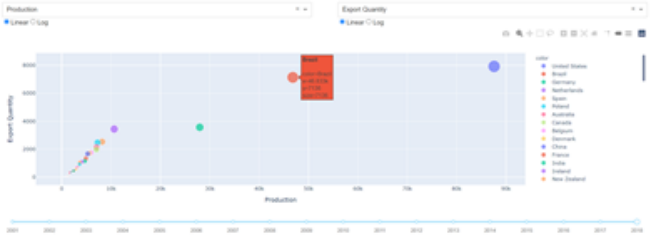

Fig. 28. Brazil one of the largest meat export

From the Treemap related to obesity percentage and total meat consumed, we can notice that **the counties with the highest meat consumed are also the countries with the highest obesity percentage**. The USA and Australia are the top countries in meat consumption (Kg\Capita\Year)

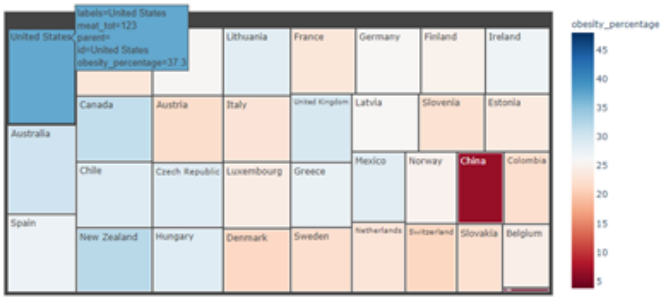

Fig. 29. Correlation between meat consumption and obesity

We can also notice how **in recent years in advanced countries, meat production is stable even if the population is increasing**. This fact may suggest a preference to import meat from low-cost countries or delocalised production to increase profit, preferring a cheaper massive quantity to meet the increasing meat demand.

In Italy, for example, the case of particular traditional meat Bresaola made from a specific local region called "Valtellina", recently has been at the centre of a scandal. The traditional meat has been replaced with Zebù meat from Argentina that, from a taste perspective, is very similar but much cheaper. [15]

Another consideration is that the most developed countries have the major demand for meat, but today also the side effects of meat consumption concern so probably in the future we will see a reduction of meat production in these countries where there is more education and information.

From the analysis conducted, **it does not seem there is a direct correlation between Covid-19 spreading and meat consumption/production.** Despite this, some more deep consideration must be done from the analysis findings. It is a truth that the **Pandemic started in the country with the highest meat production and largest meat export rate**. It is known that researchers suspect that the source of the current Covid-19 spread is the meat market in Wuhan, where many species of wild and domestic animals have been within the so-called *"Wet Markets"*. In China, since 1990, the country shifted from the primary pig meat producer to a considerable quantity of poultry and beef due maybe to a change in the wellness of the population demanding more variety of food, especially white meat.

Unfortunately, the threat to health is not limited only to these places, as already proved by the recent events. According to experts, o**ne of the main epidemiological risk factors is currently the conventional livestock system**. Particularly dangerous in this regard are intensive farming, where the vast majority of animals intended for human consumption are confined. [16] Due to the very high density and low genetic diversity of the animals raised, farms offer an ideal space for the rapid spread of viruses. This phenomenon is also favoured by the very high intensity of production, which causes chronic stress to animals and weakens the immune system. Among the potential transmission factors of zoonoses are long-distance transport and the vast supply chain used by the agri-food industry and already shown in the data analysed as meat trade.

According to the *WWF* report *"Living Planet"*, **current food production cycles (and their ever-increasing and growing consumption) are altering the planet's ecosystems and leading to the decline of wildlife**. "We are sleepy walking on the edge of the precipice," said Mike Barrett, of the WWF. «*It is as if we had emptied North America, South America, Africa, Europe, China and Oceania. These are the proportions of what we have done* ". Several scholars argue that the world has started a sixth mass extinction, the first to be caused by one species in particular: *Homo Sapiens*. [17]

C. Conclusion Marcin

One of the main conclusions from the project is the benefit of using cloud hosted databases over locally hosted deployments. Fully managed Mongo DB Atlas database and PostgresSQL provide stable systems, with simple setup, well design UIs, comprehensive documentation, backups at zero cost.

Considering the difficulties some faced with installation of DB on the Virtual Nodes - and most of all the advantage of giving access to the same database resources for all members of the team (vs lack of access for other team members in case of locally hosted DBs) - cloud-hosted solution should be considered default for future work assignments.

References

[1] Huong, N., Nga, N., Long, N., Luu, B., Latinne, A., Pruvot, M., Phuong, N., Quang, L., Hung, V., Lan, N., Hoa, N., Minh, P., Diep, N., Tung, N., Ky, V., Roberton, S., Thuy, H., Long, N., Gilbert, M., Wicker, L., Mazet, J., Johnson, C., Goldstein, T., Tremeau-Bravard, A., Ontiveros, V., Joly, D., Walzer, C., Fine, A. and Olson, S., 2021. Coronavirus testing indicates transmission risk increases along wildlife supply chains for human consumption in Viet Nam, 2013-2014.

[2] Chuluunsaikhan, T., Ryu, G., Yoo, K., Rah, H. and Nasridinov, A., 2021. Incorporating Deep Learning and News Topic Modeling for Forecasting Pork Prices: The Case of South Korea.

[3] Verma, P., Dumka, A., Bhardwaj, A., Ashok, A., Kestwal, M. and Kumar, P., 2021. A Statistical Analysis of Impact of COVID19 on the Global Economy and Stock Index Returns.

[4] Oecd-ilibrary.org. 2021. OECD iLibrary | OECD-FAO Agricultural Outlook (Edition 2020). [online] Available at: <https://www.oecd-ilibrary.org/agriculture-and-food/data/oecd-agriculture-statistics/oecd-fao-agricultural-outlook-edition-2020_4919645f-en> [Accessed 16 April 2021].

[5] Prieto, J., Malagón, R., Gomez, J. and León, E., 2021. Urban Vulnerability Assessment for Pandemic surveillance: The COVID-19 case in Bogotá, Colombia

[6] Verma, P., Dumka, A., Bhardwaj, A., Ashok, A., Kestwal, M. and Kumar, P., 2021. A Statistical Analysis of Impact of COVID19 on the Global Economy and Stock Index Returns.

[7] Rosen, L., Lim, A., Felt, J., Carrier, L., Cheever, N., Lara-Ruiz, J., Mendoza, J. and Rokkum, J., 2021. Media and technology use predicts ill-being among children, preteens and teenagers independent of the negative health impacts of exercise and eating habits.

[8] Medium. 2021. Analytics on FIFA 2019 Players!. [online] Available at: <https://towardsdatascience.com/analytics-on-fifa-2019-players-b63747958d79> [Accessed 18 April 2021].

[9] K. Pierre-Louis, "No One Is Taking Your Hamburgers. But Would It Even Be a Good Idea?," The New York Time, 2019.

[10] G. H., "https://science.sciencemag.org/content/361/6399/eaam5324.abstract," *Science,* vol. 361, no. 6399, 2018.

[11] M. Smallcombe, "The Top 4 ETL Python Frameworks," May 2020. [Online]. Available: https://www.xplenty.com/blog/top-etl-python-frameworks/.

[12] L. L. Pipino, "Data quality assessment," *ACM,* vol. 45, no. 4, 2002.

[13] K. Nishida, "Exploratory," 2019. [Online]. Available: https://exploratory.io/note/kanaugust/2617200410576325.

[14] J. H. Mustafa Zia, "Brazil Once Again Becomes the World's Largest Beef Exporter," 2019. [Online]. Available: https://www.ers.usda.gov/amber-waves/2019/july/brazil-once-again-becomes-the-world-s-largest-beef-exporter/.

[15] Ec.europa.eu. 2021. [online] Available at: <https://ec.europa.eu/smart-regulation/impact/ia_carried_out/docs/ia_2009/sec_2009_0670_en.pdf> [Accessed 2 April 2021].

[16] E. Connects, "Intensive Farming," 2013. [Online]. Available: https://www.everythingconnects.org/intensive-farming.html.

[17] T. Guardian, "Humanity has wiped out 60% of animal populations since 1970, report finds," The Guardian, 2018. [Online]. Available: https://www.theguardian.com/environment/2018/oct/30/humanity-wiped-out-animals-since-1970-major-report-finds.