An Exploration into Predicting Crop Production.

X00137773

Azees Fetuga

Contents

[Research Proposal 3](#_Toc70258943)

[Background: 3](#_Toc70258944)

[Objectives 3](#_Toc70258945)

[Prototyping & Testing 3](#_Toc70258946)

[Data Requirement: 3](#_Toc70258947)

[Risk Assessment 3](#_Toc70258948)

[Project Methodology 4](#_Toc70258949)

[Project Approach 4](#_Toc70258950)

[Project Plan: 4](#_Toc70258951)

[Research Document: 6](#_Toc70258952)

[Artical 1 : Efects of Drought on Crop Production and Cropping Areas in Texas: 6](#_Toc70258953)

[Artical 2: Prediction of Crop Yield using Regression Analysis 6](#_Toc70258954)

[Artical 3: Agricultural Crop Yield Prediction Using Artificial Agro Algorithm in Hadoop: 7](#_Toc70258955)

[Datasets Gathered: 9](#_Toc70258956)

[Dataset 1 Name: 1711 - 2016 Rainfall (Ireland) 9](#_Toc70258957)

[Dataset 2 Name: Crop Yield 1985 - 2007 and Crop Yield 2007 - 2019 9](#_Toc70258958)

[Dataset 3 Name: GNP (Ireland Billions USD) 9](#_Toc70258959)

[Dataset 4 Name: GDP (Ireland Billion USD) 10](#_Toc70258960)

[Dataset 5 Name: CO2 emissions Metric Tons 10](#_Toc70258961)

[Dataset 6 Name: Agricultural Land (sq. Km) 11](#_Toc70258963)

[Dataset 7 Name: Population Growth % 11](#_Toc70258964)

[Dataset 8 Name: Fertilizer Consumption (kilograms Per Hectare of Arable Land) 12](#_Toc70258965)

[Dataset 9 Name: Exports of cereal flour cereal, flour, starch, milk preparations and products 12](#_Toc70258966)

[Dataset 10 Name: Imports of Cereal, flour, starch, milk preparations and products 12](#_Toc70258967)

[Conclusion: 12](#_Toc70258968)

[Data Pre-processing Report 13](#_Toc70258969)

[Step 1: Importing data 13](#_Toc70258970)

[Step 2: Formatting numbers and money values and Indexing dates. 13](#_Toc70258971)

[Step 3: Plotting Values and my thoughts on each plot 13](#_Toc70258972)

[Step 4: Searching for Missing Data 14](#_Toc70258973)

[Step 5: Marking and imputing Missing Data. 14](#_Toc70258974)

[Step 6: Attribute / Feature Selection and Filters 14](#_Toc70258975)

[Step 7 (Not included Project due to poor results): Data Augmentation. 15](#_Toc70258976)

[Step 8: Normalization and Standardising 15](#_Toc70258977)

[Report on Model selection and development: 15](#_Toc70258978)

[Multiple Linear Regression Model 16](#_Toc70258979)

[Deep Nerual Network Regression Model: (Keras Regression Model) 16](#_Toc70258980)

[Random Forest Regression Model 18](#_Toc70258981)

[Decision Tree Regression Model 19](#_Toc70258982)

[Lasso Regression Model 20](#_Toc70258983)

[Ridge Regression Regression 21](#_Toc70258984)

[Report on Model Performance: 22](#_Toc70258985)

[MLRM: 22](#_Toc70258986)

[Keras Regression Model (KRM): 23](#_Toc70258987)

[Random Forest Regression Model (RFRM): 23](#_Toc70258988)

[Decision Tree Regression (DTRM): 24](#_Toc70258989)

[Lasso Regression Model (LRM): 24](#_Toc70258990)

[Ridge Regression Model (RRM) 24](#_Toc70258991)

[Generalization Performance: 25](#_Toc70258992)

[Final Thoughts on Completed Project. 26](#_Toc70258993)

[Data Augmentation: 26](#_Toc70258994)

[Datasets Gathered: 26](#_Toc70258995)

[Model Selection: 26](#_Toc70258996)

[Model Training Times: 26](#_Toc70258997)

[Model Generalization: 26](#_Toc70258998)

# Research Proposal

## Background:

As the world population continues to grow at a exponential rate so does the demand for its resources, thge most crictal resource we could have is growth as it agriculture a vital key stone which is needed to further our growth as a spieces, that is what the purpose of this project to provide a future predication of one or multiple crops using several econimic factors and climate factors as predictors.

### Objectives

1. Exploration of the predictors
2. Model Exploration
3. Model Comparison
4. Evaluation of generalizability of the best model

### Prototyping & Testing

Multiple Linear Regression Model

### Data Requirement:

The attributes of data needed for this project is

* Numerical Time Series typed data
* Data is from a sourced from a repuable source
* The quanity of data is suffice for use or for augementation
* Data relates to either the economy or climate

The proposed datasets to be used are Met eireann rainfall dataset 1711 - 2016 Rainfall ( Ireland ), Crop Yield 1985 - 2007 and Crop Yield 2007 - 2019, CO2 emissions Metric Tons, Population Growth %, Agricultural Land (sq. Km), : Fertilizer Consumption ( kilograms Per Hectare Of Arable Land ), Temp Grid

### Risk Assessment

The risk involved in the project are the following

* The sample size of the data may not suffice to make accurate predications or good models.
* The scope of the project may be swayed during the duration of it.
* Models and Predications that suffice may not be achieved

## Project Methodology

### Project Approach

* Identify user/market needs
* Analysis of user needs and solutions
* Design a solution or test available solutions
* Implement and deploy the solution(s) in order to meet requirements
* Test (and/ or compare) the solution(s) – does it meet requirements and objectives?
* Modify design and/ or solution then test again for generalizability.
* Report findings.

### Project Plan:

#### Project Phases

1. Phase 1 Research:
2. Phase 2 Development:
3. Phase 3 Outcomes and Presentation:

#### Deliverables and Milestones

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Task | Start | End | Duration |
|  |  |  |  |
| Phase 1 Research |  |  |  |
| Task 1.1 Idea | 12 October 2020 | 26 October 2020 | 14 |
| Task 1.2 Final Idea | 26 October 2020 | 02 November 2020 | 7 |
| Task 1.3 Research Draft | 02 November 2020 | 09 November 2020 | 7 |
| Task 1.4 Research | 09 November 2020 | 07 December 2020 | 28 |
| Phase 2 Development |  |  |  |
| Task 2.1 Data Set Selection and Preparation | 07 December 2020 | 25 January 2021 | 49 |
| Task 2.2 Data Pre-processing | 25 January 2021 | 01 March 2021 | 35 |
| Task 2.3 Model Development | 01 March 2021 | 22 March 2021 | 21 |
| Phase 3 Outcomes and Presentations |  |  |  |
| Task 3.1 Performance and Model outcome | 22 March 2021 | 12 April 2021 | 21 |
| Task 3.2 Final Documents and Uploads | 12 April 2021 | 26 April 2021 | 14 |
| Task 3.3 Presentations | 26 April 2021 | 26 April 2021 | 0 |

#### Gantt ChartChart Description automatically generated

# Research Document:

Examinination of 5 different articals in order to gain insight into the approaches and techiques that have been conducted in the area regrading my projects from those articals heres what i found

## Artical 1 : Efects of Drought on Crop Production and Cropping Areas in Texas:

In this artical the author explores the effect of a drought on the crop and cropping areas in texas.

The author starts by gathering data from sources source as CDL Cropland Data Layer, USDA National Agricultural Statistics Service, USGS elevation

And then states there is a need to for more research on the effects of climate change on crop productions due to drought

#### Step 1: CDL(Cropland Data Layers) analyze

He then used tools such as ArcGis Gis (A geographic infromation system) to analyze the CDL in the study area and yield data as well for the four major crops in the state.

#### Step 2: Spatial Analyst

Using ARCGIS Spatial analyst Tool “Extraction” to extract the cropping area from each climate zone.

#### Step 3: Impact Analysis

#### 4 Major croups during the periods

They analysed the impact of drought on crop yields and cropping area from each climate zone before, during and after severe drought for the 2011- 2013 period in three different periods 2008 to 2011, 2011 to 2013, and 2013 to 2016 and then compare the deviations from the baseline data (actual crop production).

Step 4: Irrigation mitigation Assement:

Comprasion between the between irragated and rainfed crop yield data

## Artical 2: Prediction of Crop Yield using Regression Analysis

The paper attempts to predict Crop yeild using regression analysis,

#### Stage 1: Compute a Linear Regression Model

In this first stage the author explains how to create a successful model the following conditions are needed

1. Linearity, 2. Nearly normal residuals and 3. Constant variability.

“Linearity explains that there should be a linear relationship between the response variable and the explanatory variable of the model”.

“This condition can be easily checked for using a histogram or a normal probability plot of the residuals. If the histogram happens to be symmetric then we can interpret that the residuals are normally distributed. In case of the plot of residuals”

#### Stage 2: Compute the Residual Values:

In the stage the author finds the difference between the predicted values and the actual results this know as the residual value.

“Residual = Observed value – Predicted value e = y – ŷ”

#### Stage 3: Compute the Residual Sum of Squares and Obtain the R2

In this stage the author finds the Residual Sum of Squrares and the R2

The author states the Residual sum of squares can be reffered to as discrepancy between the data and the estimation model and also that strength of any linear model is generally evaluated using R2 .

“The value of the R2 is always between 0 and 1 that corresponds to the variability of the response variable that is explained by the model.”

#### Stage 4: Implementation

In this stage the author impements stage 1 -3 using data from the given data table i.e rice production, matlab environment used for linear regression

## Artical 3: Agricultural Crop Yield Prediction Using Artificial Agro Algorithm in Hadoop:

In the artical the paper proposes how to make agriculture well organized by predicting and thus improve the crop yields by using soil information.

The paper suggests 3 steps to be taken when try to predict crop production Assessing Variability, Managinging variability and Evaluation before moving to their

Assessing variability:

In order to assess the variability the paper suggest to do the following

* Selection of important soil/ plant factors
* The Identification of a quick method of measuring soil properties
* The identification of important soil factors affecting yield variation

Managing Variability:

In order ot get more accurate re

* Quality seeds should be selected
* Humidity management of soil
* Accurate sowing of seeds
* Fertilizer and Pesticides requirement estimation

Evaluation:

* Verification of farm/farmer field
* Comparison of yield performance and economic analysis of precision and traditional agriculture

The proposed architecture of the paper was:

1. Get Multiple Agriculture Data sets:

* Obtained from Hadoop Distributed File System

1. Analysing the datasets

* Analysis Algorithm:

Varval,total.avg

For each val then

Begin

ReadFile()

End

For each total then

Begin

Total+=val

End

Avg=total/N

1. Classification:

* Prediction Algothorim :

Varmean,sum

Var pi= 1068966896 / 340262731.

Get NormalizedData from DFS

For each row then Begin Sum+=value

End

Return sqrt((sumSq-((sum\*sum)/n))/(n-1))

1. Predict for the Crop and soil to be planted.

* Prediction Algorithm:

Var like, what,kclass

Get ClassifiedData from DFS

If like then

Begin

Likelihood(like,kclass)

End

Function likelihood(like,kclass)

Begin

Case CROP:

putInCrop(like);

break;

Case SOIL :

putInSoil(like);

break;

Case DISEASE:

# Datasets Gathered:

The datasets gathered were all selected using the data requirements stated in the research proposal .

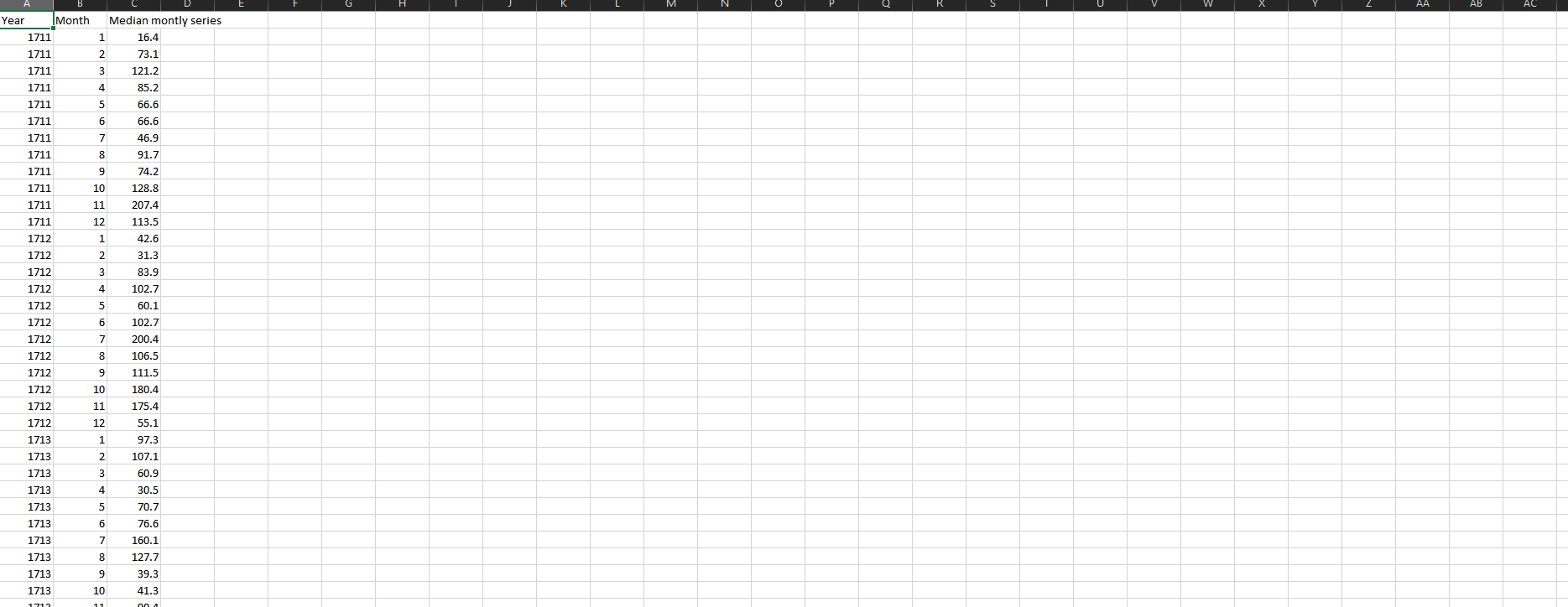
### Dataset 1 Name: 1711 - 2016 Rainfall (Ireland)

Place Acquired: <https://www.met.ie/>

Relevance to Project Goal: Key factor in crop production as well plants need water to grow, differences in rainfall will affect growth of certain crops.

Data Type: Time series monthly

Description of Work done with csv: python script made to average the rain fall for each year.

Image of csv with factors: 

### Dataset 2 Name: Crop Yield 1985 - 2007 and Crop Yield 2007 - 2019

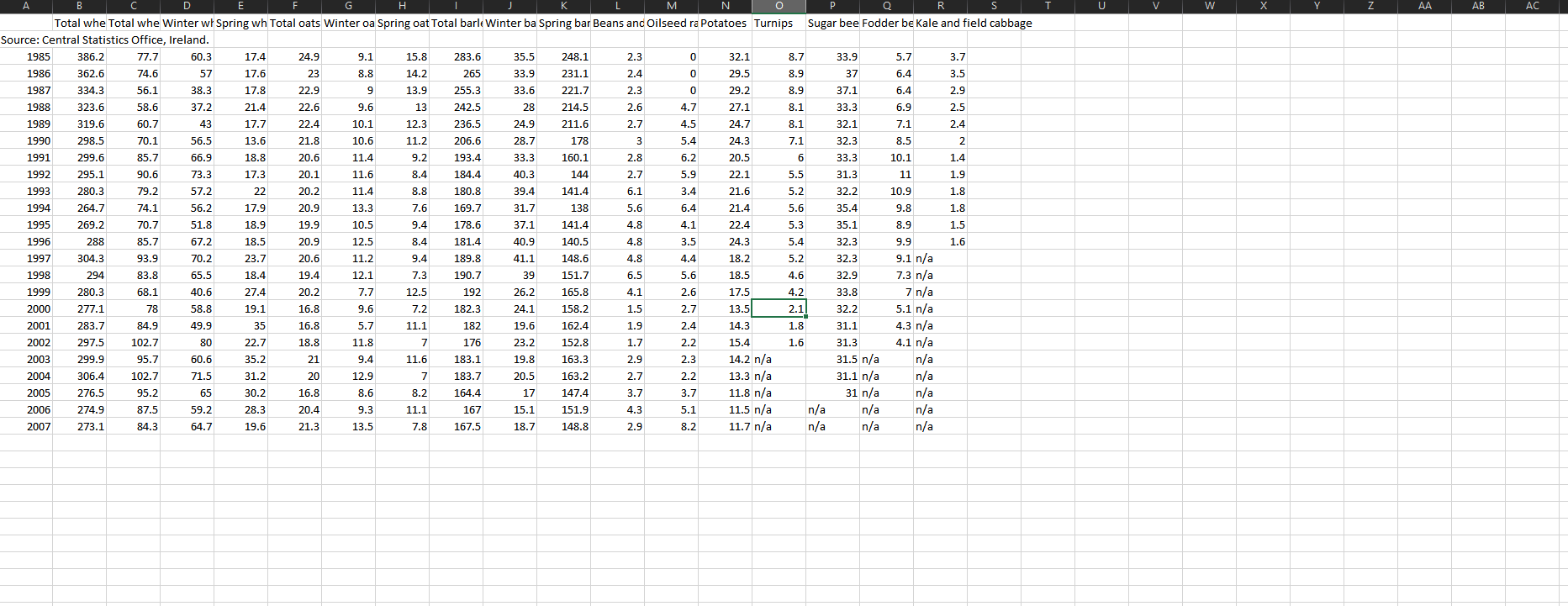
Place Acquired: https://data.cso.ie/

Relevance to Project Goal: contains a different crop and their associated yield number per year (Selected crop to examine (potatoes))

Data Type: Time series yearly

Description of Work done with csv: Combined the two datasets into one.

Image of csv with factors:



### Dataset 3 Name: GNP (Ireland Billions USD)

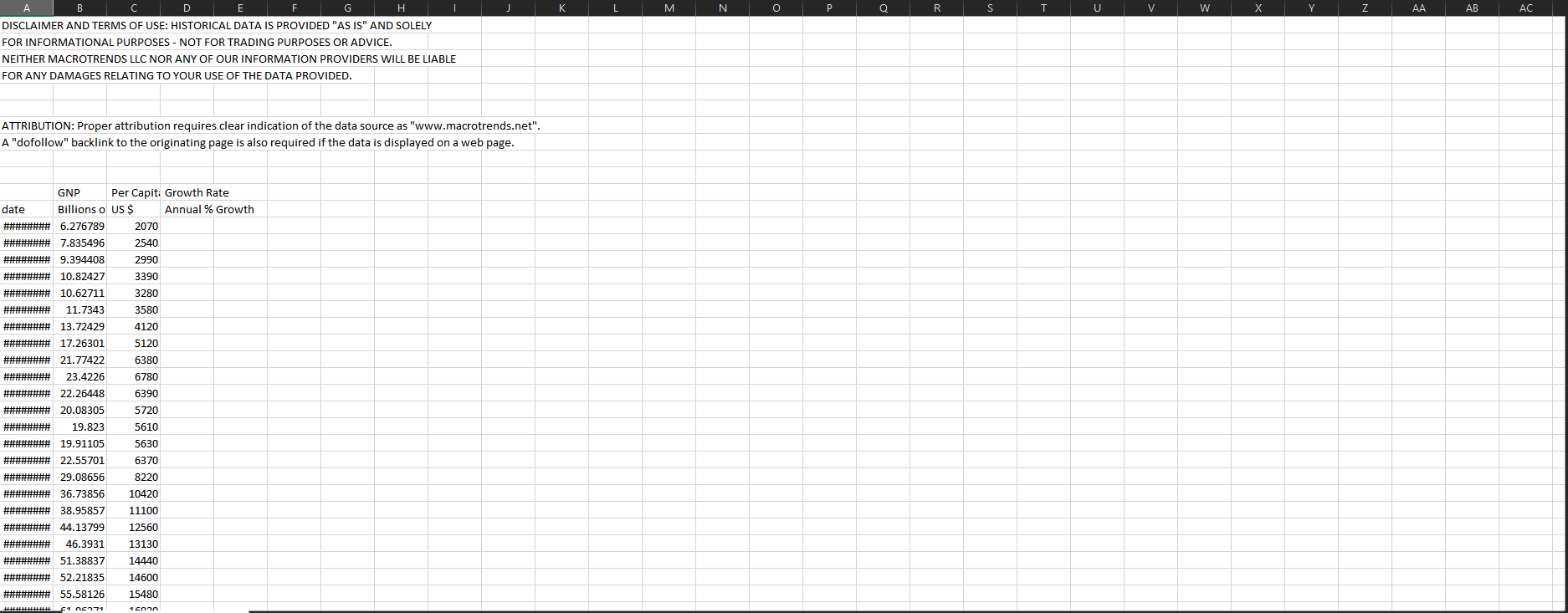
Place Acquired: https://www.macrotrends.net/

Relevance to Project Goal: Looking to see if there is a dependency between gdp and crop production.

Data Type: Time series yearly

Description of Work done with csv: None.

Image of csv with factors:



### 

### Dataset 4 Name: GDP (Ireland Billion USD)

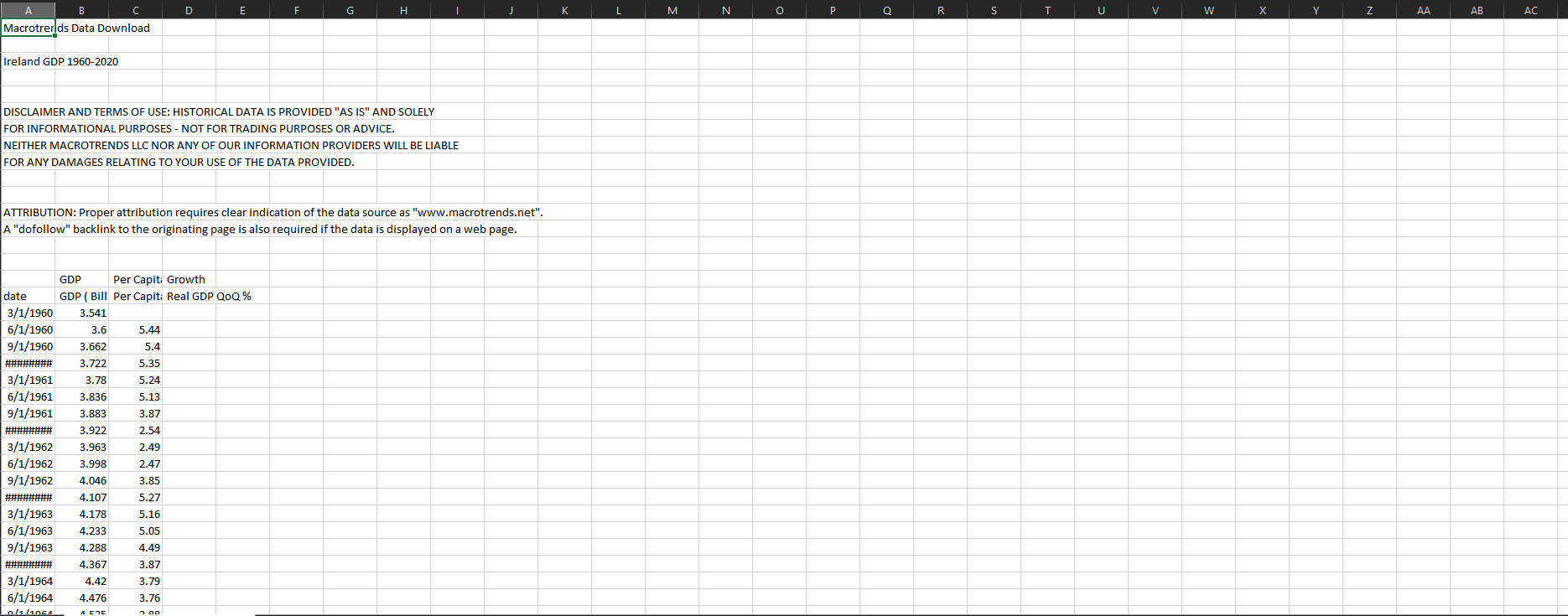
Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: Looking to see if there is a dependency between GDP and crop production.

Data Type: Time series yearly

Description of Work done with csv: None.

Image of csv with factors:



### 

### Dataset 5 Name: CO2 emissions Metric Tons

Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: Key factor in crop growth,

Data Type: Time-series yearly

Description of Work done with csv: None.

Image of csv with factors:

### 

### Dataset 6 Name: Agricultural Land (sq. Km)

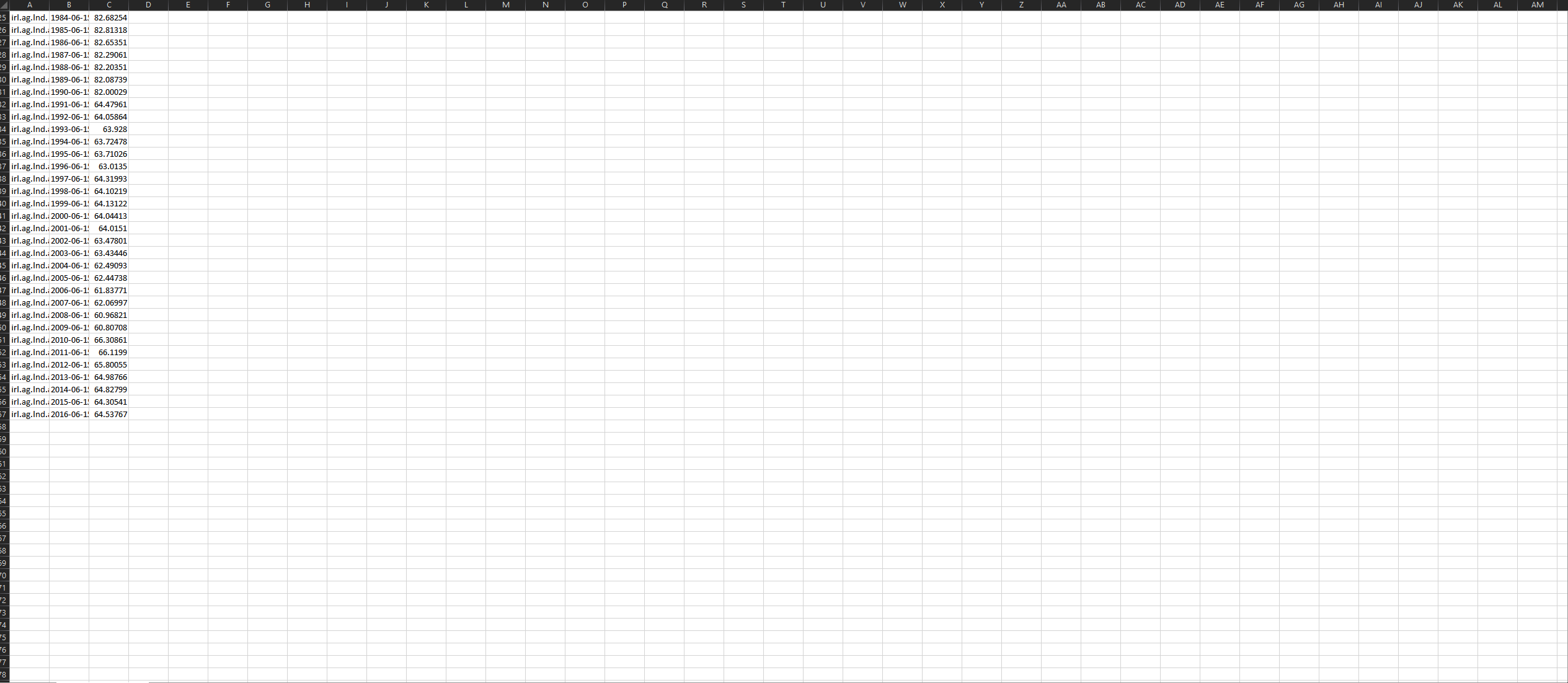
Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: Land available for crop production

Data Type: Time-series yearly

Description of Work done with csv: None.

Image of csv with factors:

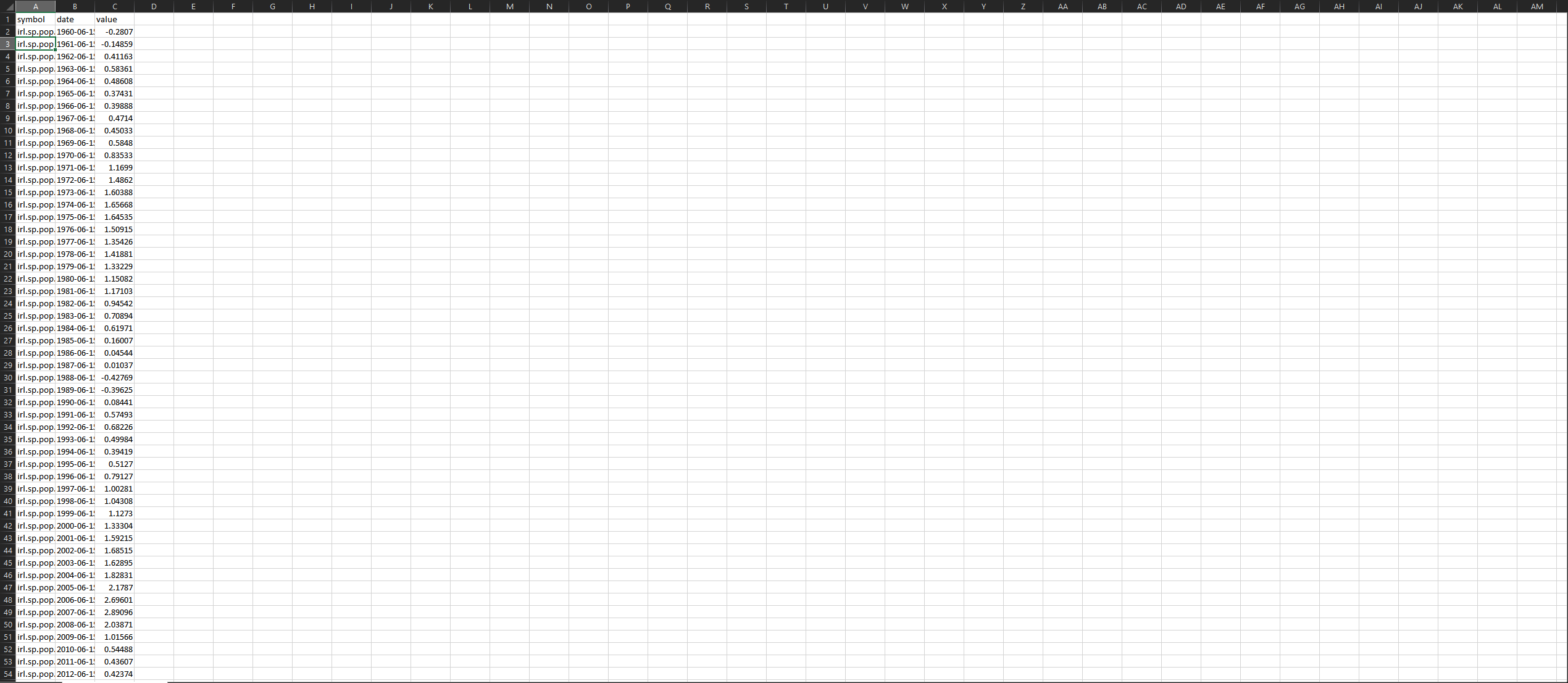


### Dataset 7 Name: Population Growth %

Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: Looking for a dependency between the two.

Data Type: Time-series yearly

Image of csv with factors: 

### Dataset 8 Name: Fertilizer Consumption (kilograms Per Hectare of Arable Land)

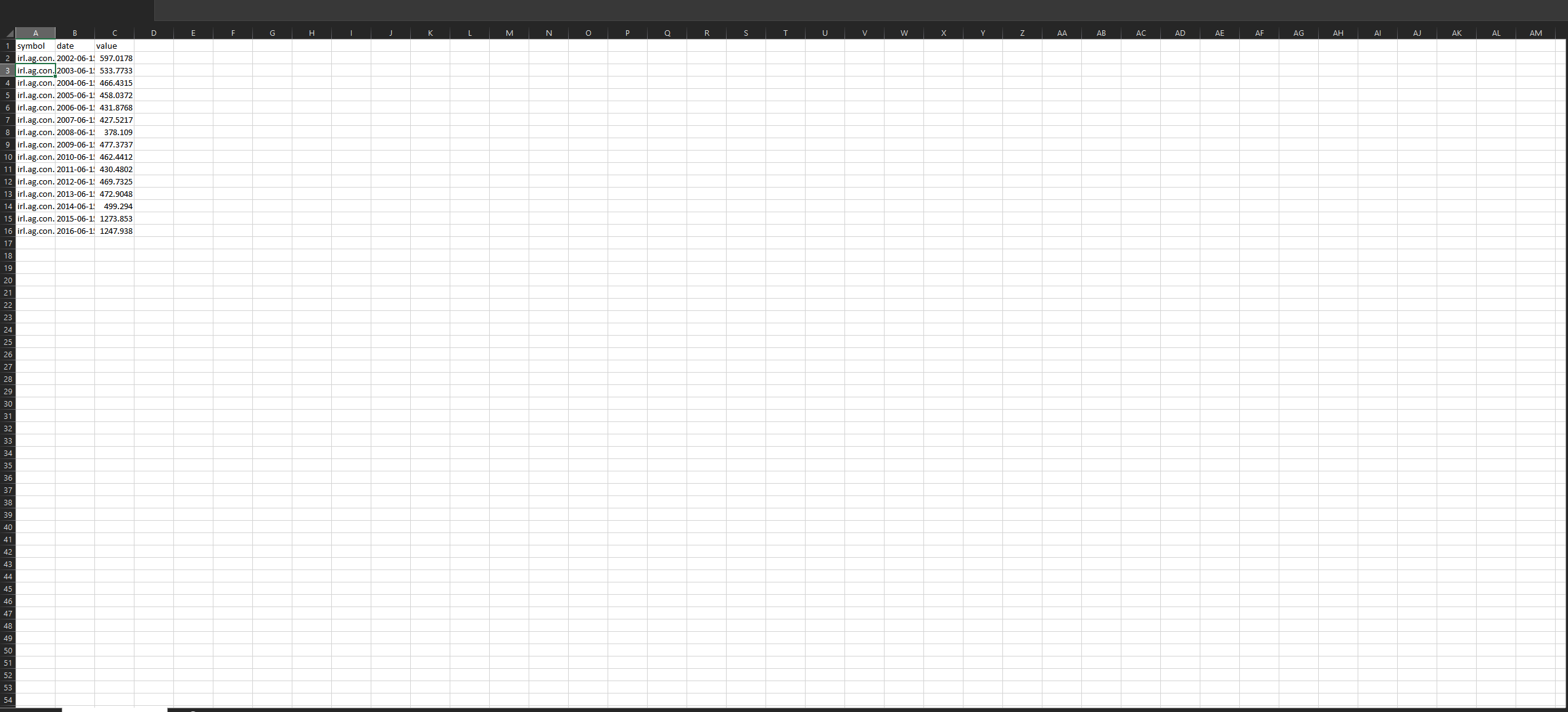
Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: The amount of fertilizer used impacts crop yeild quanties.

Data Type: Time series yearly

Description of Work done with csv: None.

Image of csv with factors:



### Dataset 9 Name: Exports of cereal flour cereal, flour, starch, milk preparations and products

Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: The amount of exports influences crop production

Data Type: Time series yearly

Description of Work done with csv: None.

Image of csv with factors:

### Dataset 10 Name: Imports of Cereal, flour, starch, milk preparations and products

Place Acquired: https://tradingeconomics.com/

Relevance to Project Goal: The amount of imports influences crop production

Data Type: Time series yearly

Description of Work done with csv: None.

Image of csv with factors:

## Conclusion:

For this project I gather data from the various sources on the internet listed above.

The criteria for selecting my datasets were as follows.

· Are they of relevance to crop production.

· Do they add scope to the goal of my project.

· Is the dataset in a useable format.

· Does the dataset have enough information to be valuable.

· Are values in the dataset accurate

· Does the dataset compliment other datasets already acquired.

· Must be geolocated to Ireland.

· From a reliable source

Most of the datasets were not changed to get into a useable format apart from two datasets in which I used a python script to average the rainfall for each year.

There was no need to do ethical work as all the datasets were acquired from reliable sources as mentioned above.

Data from before 1985 was filter out of all datasets to keep the data pool as modern as possible as well as converting all.

# Data Pre-processing Report

The programming language used fort this throughout was Python as it the language I have the most familiarity with and am comfortable using.

### Step 1: Importing data

Firstly, I imported the necessary libraries to be used through (pandas, NumPy, matplotlib, datetime and sklearn) as well as the clustered dataset in a data frame into a Jupiter notebook file.

### Step 2: Formatting numbers and money values and Indexing dates.

To make the data into a more readable format, applied a function that rounds a number to the closest 1 decimal point to all columns with numerical values as well indexing the date column in the data frame.

### Step 3: Plotting Values and my thoughts on each plot

All columns in the data frame were plotted as a time series plot using a function, I made that plots all the columns in a data frame, I later then examined each plot to identify any outliers or any noteworthy trends/fluctuations.

Crop Production Plot (Irish Potatoes): The plot shows a depreciating as the number of potatoes being produced national dwindling throughout the entire of timeframe with no noteworthy spikes.

Irish GDN: The plot is a positive one, showing a noteworthy explosive growth through the years 2002-2008 and is the met by a small decline (Cause by the recession) before resuming it explosive growth for reminder of the years.

Irish GDP: Very similar to GDN which is to be expected.

Population Change: Shows that the population % increase has been rapidly before a very harsh fall around from 2007 -2011 and is now rapidly regaining the same momentum it had earlier in the timeframe.

Co2 Emissions: Shows Bell curved shaped plot with the peak being in the early 2000.

Agricultural Land: Very Sudden Drop a from Mid 80 km2 to mid-60 in the early 90s my speculation would a significate event occurred around that time frame that led to this sharp drop.

Yearly Rain: Wildly flucating from year to year

Exports and Import: Both plots are quite similar, both show an aggressive growth across the timeframe while tapering off through the 2010s.

Fert Consumption: Stead y line of usage before a massive surge from to 2014 -2016

### Step 4: Searching for Missing Data

Ran various in-built function to search through the data frame to find any data that was missing or marked as N/A, Data frame was missing 43 variables across all columns:

### Step 5: Marking and imputing Missing Data.

Used in-built functions to mark all missing data as NAN before running another in-built function to remove impute them from the data frame.

### Step 6: Attribute / Feature Selection and Filters

Various elements were considered and evaluated when selecting the features which were to be used.

1. Correlation between the selected crops.

From running a correlation analysis, I can determine for each variable the following.

* Irish GDN: The correlation between this and the selected crops is very strong and negative with a score of -0.9 indicating to that two are when one of the values goes up the other must fall (High Negative correlation)
* Irish GDP: Similarly, to GDN this attribute as closely matches with its high negative correlation.
* Population Change: with a correlation -0.29 there is some relation with the selected crop and this variable albeit weak,
* Agricultural Land: The correlation between this came is very low and negative with a score than 0.1.
* Yearly Rain: The correlation score was also very low and negative with score lower than 0.1.
* Exports: A Strong negative correlation was found with score of over 0.8
* Import: Like Exports shows a high negative correlation

2. Variance:

The variance for all the attributes in the data frame met my expectations as they high scores indicating a good fluctuation of change in each attribute.

3. Relationship with other variables:

When correlating all attributes, I found the following for each attribute:

* GNP: Strong positive relations with GDP, Imports and Exports
* Weak correlation relations with the rest of the attributes
* GDP: Like GNP although has higher correlation with CO2 and Population.
* Population Change: Shows moderate correlations between CO2 and Agricultural Land
* Weakly Correlated with the rest of the attributes
* Very weakly correlated to Exports and Imports > 01.
* CO2: Shows Moderate Correlations with Population Change, Imports and Exports
* Weakly correlated to the rest of the attributes
* Agricultural Land: Weakly correlated with everything expect population change and CO2 Emission.
* Yearly Rainfall: Weakly Correlated to everything
* Exports: Highly Correlated with GDP, GNP, and imports
* Moderate Correlations with C02 emissions
* Imports: High correlation to Exports

### Step 7 (Not included Project due to poor results): Data Augmentation.

As my dataset was quite limited in size, I attempted to artificially create new instances to use to train my models better using a data augmentation technique known as

In which random instances are created within the range of the of standard deviation of the data.

### Step 8: Normalization and Standardising

I choose to standardize my data as through examining the visual representations of each attribute the I cannot make assumptions about the distributions of the data and to scale,

Using in built functions I first split the data into arrays X and Y before applying a scaler to X and then then converting them back to data frame before finally concatenate X and Y into a single data frame.

# Report on Model selection and development:

The dataset was s split, where 40% of it is used for testing and the reminder for training.

x\_train, x\_test, y\_train, y\_test = train\_test\_split (X, y, test\_size=0.4,random\_state=1)

The models selected and developed upon are the following:

### Multiple Linear Regression Model

Multiple linear regression is attempts to model the relationship between one or more explantory variables and the response variable by fitting a linear equation to the observed data in order to predict the response variable or explore the effect of a explantory varibale

Equation for multiple linear: Y=a+b1X1+b2X2+b3X3+…. +bnXn+u

Tools and infastruture used:



I chose to use the criteria of Mean Square Error (MSE), Root Mean Square Error (RMSE) as my Selection criteria for all models created as it shows how close the model a models estsimates for the target outcome were which I feel is crucial for determining which model is objectively the best.

Created evaluation function: 

### Deep Nerual Network Regression Model: (Keras Regression Model)

This is Nerual network that is capable to predicting numbers. In

my case I built regression model using the deep learning framework Keras which is an neural network library which provides high level API which is run ontop of Tensorflow, CNTK and Theano librabries

For the development of this model I create two sequential models ,a intial model with no paramaters at all where i explored the topology to find best the model for generalizability found the best network topology and a hyper tuning model using the same topology as the inital model using several different paramater combations and cross validated in order to find the best weights.A bias was introduce to the first and second layer.

The initial model topology network settled on was the following:



As I found this set of parameters to produce the most optimal mse , rmse and r2 scores and for generalizability thorughout the experiments I conducted.

In order to hyper tune the model a grid search and random search was conducted with the following parameters to be tested:



The parameters expolored to find the optimal weights were the epoch, optimizer, init and batch\_size.



A bias was used on the first 2 layers of both models as to try and tackle the high amount of multicolinterity found between the predictors which proved to help reduced the overfitting.

### Random Forest Regression Model

Random Forest is a Supervised Learning algorithm which uses ensemble learning methods (meaning it combines several base models together to produce one optimal predictive model) for both classification and regression. Random Forest fits multiple CART trees to independent “bootstrap” samples of data and then aggregating the predictions.



rf\_random = RandomizedSearchCV (estimator = rf, param\_distributions = random\_grid, n\_iter = 100, cv = 4, verbose=2, random\_state=1,)

grid\_search = GridSearchCV (estimator = rf, param\_grid = param\_grid,

                          cv = 4, verbose = 2)

The parameters test in both Grid Search and Random Search were the following:

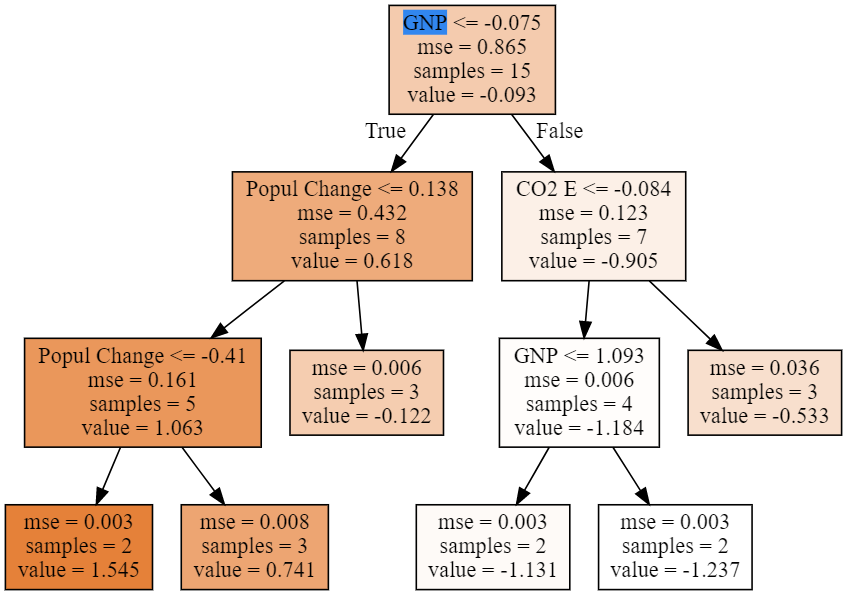


### Decision Tree Regression Model

A Decison Tree is a predictive model that uses a set of binary rules to calcualate the desired value, each of unique tree is a simple model that has branches and nodes and leaves.







### Lasso Regression Model

Lasso regression is a type of linear regression model that uses shrinkage (A technique where data values are shrunk towards the mean or a central point)

I chose to use this model as part of this project as in the earlier stages of the project i dicovered a high amount of multicolineratity within my predictors as an attempt to avoid the overfitting this will cause lasso was my choice as it will preform regularization on the predictors in question thus resolving this by using L1 regularization.

Both Grid Search and Random Search was used to for this model to compare method of hypertuning which gave better results.

Testing of which penatly (“alpha”) to determine how many features are important were the only parameter in this model although this still may run the risk of overfitting as it will only tell me if a feature is responsive to the model not the data itself.



### Ridge Regression Regression

I chose to use this model as part of this project as in the earlier stages of the project I dicovered a high amount of multicolineratity within my predictors as an attempt to avoid the overfitting this will cause was my choice as it will preform regularization on the predictors in question thus resolving this by using L2 regularization. I feel this will produce more accurate estimates than Lasso Regression but as my target variable has many predicators with coefficients of about the same sizei felt the was the ideal choice of Regression to carry out.

Both Grid Search and Random Search was used to for this model to comapre which gave better results

Again only alphas is the only parameter for this model.



# Report on Model Performance:

The various models were all evaluated using metrics such as Mean Square Error (MSE), Root Mean Squared Error (RMSE) and R Square Score (R2^) and a Residual plot, all models except for the MLRM model were Cross-Validated a total of 4 times.

R-Squared: Is the measure of how close the data is fitted to the regression line measure between the values of 0-1 or 0% - 100% but is not a complete indicator of how good a model.

Mean Squared Error: Used to measure the difference (error) between the actual result and the predicted value.

Root Mean Square Error: Used to measure how concentrated the data is around the line of best fit.

### MLRM:

MSE 0.2705780212257958 RMSE: 0.5201711460911647 R^2 SCORE: 0.7688724482973991

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | y | **R-squared:** | 0.913 |
| **Model:** | OLS | **Adj. R-squared:** | 0.796 |
| **Method:** | Least Squares | **F-statistic:** | 7.828 |
| **Date:** | Thu, 15 Apr 2021 | **Prob (F-statistic):** | 0.0109 |
| **Time:** | 16:12:03 | **Log-Likelihood:** | -2.9474 |
| **No. Observations:** | 15 | **AIC:** | 23.89 |
| **Df Residuals:** | 6 | **BIC:** | 30.27 |
| **Df Model:** | 8 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 0.1257 | 0.170 | 0.741 | 0.487 | -0.289 | 0.541 |
| **x1** | 0.3142 | 0.950 | 0.331 | 0.752 | -2.011 | 2.640 |
| **x2** | -1.0769 | 1.130 | -0.953 | 0.378 | -3.843 | 1.689 |
| **x3** | -0.5287 | 0.687 | -0.769 | 0.471 | -2.210 | 1.153 |
| **x4** | 0.0827 | 0.565 | 0.147 | 0.888 | -1.299 | 1.464 |
| **x5** | -0.7893 | 0.494 | -1.597 | 0.161 | -1.998 | 0.420 |
| **x6** | -0.1524 | 0.194 | -0.787 | 0.461 | -0.626 | 0.321 |
| **x7** | -0.3233 | 0.500 | -0.646 | 0.542 | -1.548 | 0.901 |
| **x8** | 0.4362 | 0.955 | 0.457 | 0.664 | -1.900 | 2.772 |

### Keras Regression Model (KRM):

Initial Model Performance:

MSE 0.13852274700050618 RMSE: 0.372186441183053 R^2 SCORE: 0.8816739688452808

Tuned Model Performance (Grid Search):

MSE 0.1464224075260167 RMSE: 0.382651809777527 R^2 SCORE: 0.8749260844891479

Best estimator:

{'batch\_size': 240, 'epochs': 400, 'initIn': 'normal', 'optimizerIn': 'adam'}

Tuned Model Performance using (Random Search):

MSE 0.13966470500071887 RMSE: 0.3737174132960877 R^2 SCORE: 0.8806985091403844

Best parameter:

{'optimizerIn': 'sigmoid', 'initIn': 'normal', 'epochs': 300, 'batch\_size': 60}

### Random Forest Regression Model (RFRM):

Initial Model Performance:

MSE 0.1753983097736070 RMSE: 0.41880581392049354 R^2 SCORE: 0.8501748895675518

Tuned Model Performance using (Grid Search):

MSE 0.1858224509264479 RMSE: 0.43107128288306085 R^2 SCORE: 0.8412705956698302

Tuned Model Performance using (Random Search):

MSE 0.1773658982632116 RMSE: 0.42114830910643775 R^2 SCORE: 0.8484941768906673

Best parameter:

{'n\_estimators': 311, 'min\_samples\_split': 2, 'min\_samples\_leaf': 1, 'max\_features': 'sqrt', 'max\_depth': 110, 'criterion': 'mse', 'bootstrap': True}

### Decision Tree Regression (DTRM):

Initial Model Performance:

MSE 0.09470147821771316 RMSE: 0.3077360528402760 R^2 SCORE: 0.9191060652158006

Tuned Model Performance using (Grid Search):

MSE 0.13385672850834243 RMSE: 0.3658643580732379 R^2 SCORE: 0.8856596784953373

Tuned Model Performance using (Random Search):

MSE 0.22137918895679315 RMSE: 0.4705094993268394 R^2 SCORE: 0.8108980555416486

### Lasso Regression Model (LRM):

Tuned Model Performance using (Grid Search):

MSE 0.17899195592507508 RMSE: 0.4230744094424467 R^2 SCORE: 0.8224511224428661

Best parameter:

{'alpha': 0.01}

Tuned Model Performance using (Random Search):

MSE 0.11736977115450711 RMSE: 0.34259271906231037 R^2 SCORE: 0.883576493592017

Best parameter:

{'alpha': 0.10609699216599455}

### Ridge Regression Model (RRM)

Tuned Model Performance using (Grid Search):

MSE 0.09483449365813101 RMSE: 0.3079520963691122 R^2 SCORE: 0.9059300859880627

Best parameter:

{'alpha': 1}

Tuned Model Performance using (Random Search):

MSE 0.09528632461092322 RMSE: 0.3086848305487706 R^2 SCORE: 0.9054818978105593

Best parameter:

{'alpha': 0.567297771334341}

As we can see by the results the best performing models overall respectively were

* RRM
* DTRM
* KRM

But overall across all models were quite adequate.

## Generalization Performance:

To generalize the best performing models for its intended purpose I introduced a new dataset (This Dataset was also standardized) from a different country (The United Kingdom) and used all the models to attempt to predict the output potato, the results of this are the following.

Actual Values:

[ 0.21441799 -0.26187735 0.49520345 1.32092069 0.61375731 1.14204995 -2.51232082 -0.15788273 0.31217293 -0.35963229 -0.80680913]

Multiple Linear Model:

[-1.4898228 -2.26866341 -2.08868627 1.63998003 1.72785308 1.14040926 0.09235743 -0.6125508 -0.85005919 0.57809145 1.25924085]

MSE 2.585533416796378 RMSE: 1.6079593952573485 R^2 SCORE: -1.585533416796379

Ridge Model:

[-0.90226141 -1.77122998 -1.6687798 1.55451287 1.31169188 0.63370917 0.2781069 -0.19430159 -0.81225277 0.13593352 0.81848042]

MSE 1.9043025405286504 RMSE: 1.3799646881455518 R^2 SCORE: -0.9043025405286509

Lasso Model:

[-1.19956958 -2.06999078 -1.58486394 1.96709154 1.41486014 0.41820148 0.12886455 -0.27545616 -0.88236418 0.37578589 1.17720324]

MSE 2.188392912848907 RMSE: 1.479321774614606 R^2 SCORE: -1.1883929128489075

Keras Model:

[-0.475877 -0.84611464 -0.8982895 1.2496172 0.6705711 -0.23062007 1.0678028 0.5581071 -0.96788955 -0.7080458 0.19766824]

MSE 1.8864591861340272 RMSE: 1.3734843232210652 R^2 SCORE: -0.8864591861340279

Random Forest Model:

[-0.33299787 -0.59387361 -0.62732797 0.74443734 1.0203334 -0.17453344 -0.09185338 -0.37389412 -0.94077484 -0.71164661 -0.35361668]

MSE 1.0641355017355159 RMSE: 1.031569436216252 R^2 SCORE: -0.06413550173551608

Decison Tree Model:

[ 1.54450861 -0.53281371 -0.53281371 0.74118294 -0.12230368 -0.12230368 1.54450861 1.54450861 -1.13088436 -1.13088436 -1.13088436]

MSE 2.501282606707136 RMSE: 1.5815443739292097 R^2 SCORE: -1.5012826067071368

From looking at result I can assume the data used to test the generelizablity of the models was not ideal as it diffes greatly from the data set used to train the models in terms of the impact the predicators have on the target variable, with negative r2 scores (which I will equate to a score of zero as it is the lowest r2 score i can theorically have) aswell as the massive deviations between the predicated values and the actual values leads me to beleive this.

# Final Thoughts on Completed Project.

### Data Augmentation:

In order to eliminate the constraint of size as my dataset was quite small, I attempted to conduct data augmentation on the existing data using a technique known as “Uniform Random Generation”, but my inexperience and oversight resulted in the wrong choice of technique as realised later in project, The effects of this mistake were poor results across the board for all the models built and evaluated.

### Datasets Gathered:

Gathering data relating to or of significance to the goal of the project was quite tedious as a lot of data that would have proven to be substantial were hidden behind paywalls or just lacked the requirements needed for me to include it in the project.

### Model Selection:

For the most part I feel that the models I selected to use throughout the project were appropriate to problem I was trying to solve though I if I were to redo the project, I would attempt to do an elastic net regression model as it improves upon the flaws of

### Model Training Times:

I conducted the entirety of the project using what I would consider mid-ranged hardware. Models such as Keras, Random Forest and Decision Tree took a considerable amount of time to train. Though with Random Forest the parameter n\_jobs allowed me to hasten this process.

### Model Generalization:

Limited Data meant that I had to use a dataset of a different distribution which was not ideal when trying to determine the generalizability of my models, in retrospect perhaps the model should a been fitted to a country where the data needed was more easily and readily available.