# Data Prep

## Question 1

During the exploratory data analysis phase, the dataset was read into the Jupyter notebook using the Pandas read\_excel function. Next the dataset was split into two variables with each variable being a dataset. The first variable held all of the data imported from the excel spreadsheet while the second variable contained data for 2006 only. The rows containing the 2006 data were first identified by viewing the spreadsheet and then the using drop function with the range function inserted into its parentheses with the values of where to start dropping rows from and where to stop dropping rows. These passed values to the range function encompassed all rows containing data belonging to the years 2007 to the end of 2023.

Using the default head function allowed for the first five to be viewed within the Jupyter notebook while the default tail function allowed for the last five rows to be viewed. Using these two functions was a quick and easy way to check that no rows of data were lost. A more complete check could be accomplished by using the shape function on the variables, using this command would print the total number of rows and columns to screen. With the total numbers of rows and columns matching that of the spreadsheet it could be said with confidence that no data was lost when reading in the file or when dropping rows.

In order to get an idea of the data types being used the .dtypes function was used to print to screen that data type for each column. When using this function, the returned values where of the types object, int64 and float64. The objects are text/strings and cannot be used for mathematical calculations as they are not numeric values. The int64 data types are integers/whole numbers that are 64bits or 8bytes in memory. The float64 data types are numbers with a decimal point and are also 64bits or 8bytes in memory. Both the int64 and float64 data types allow for mathematical calculations to be performed on the as they are numeric values.

Using the drop function again columns holding text values could be dropped and they had no use for performing calculations. This time when using the drop function the column names where inserted into the drop function’s parentheses and the column name values where also surround by square brackets. The drop function was also given the parameter of axis=1 so, the computer would know to drop columns. At this point to ensure the correct columns had been dropped the head function was called to check that the number of columns within the data set had in fact been dropped from six to four. This net step is not essential but was demonstrated to show it could be done. Renaming the column heading ‘Sex’ to ‘Gender’ was done using the rename function. The rename function too the parameter columns= and then the values were surrounded with curly brackets. Again, it was not essential but it did remove any opportunity for immature jokes to be made.

Next the duplicated function was used to check for duplicate rows. The result of this function when combined with the shape function prints to the screen the number of duplicate rows. Before dropping the duplicate rows its good practice to first use the count function which will return the number of rows for each column with the dataset. The drop\_duplicates function can the be called which will drop all duplicate rows found within the dataset. The count function is then called and its returned value is compared with the result of count function previously used before dropping the duplicate values. There were no duplicate values found within the spreadsheet provided by the CSO but the steps described in this paragraph are essential when preparing data for data analytics.

With the duplicates removed the next step for preparing the data was to remove all null values as they are not needed. Removing the null values ensures analysis on the dataset is based on reliable data and was achieved by first using the isnull function combined with the sum function on the data frame. These two functions combined returned the total number of null values found within the dataset. The dropna function was used to drop the null values. A count was taken after the null values had been dropped and compared with the result of the number of null values found when using the isnull function. Like the duplicate values there were no null values found within the spreadsheet provided by the CSO but the steps described in this paragraph are essential when preparing data for data analytics.

“An outlier is a datapoint or set of datapoints that are vastly different from other data points in your dataset. Sometimes they can be very high or very low.”[1]

In order to detect and remove outliers, the interquartile range scoring techniques was used. These outliers can be visually seen in boxplots and removing them increases the accuracy of the data. For example, if my dataset has 100 values that range from one to ninety-nine and the last datapoint has a value of one million this single datapoint would skew the total dataset and the resulting graphical displays would also illustrate the skew as they would try to fit the single datapoint with the value of one million into the dataset. The code for running the IQR and the code for removing outliers are in the Jupyter notebook but were not needed as the data collected by the CSO have similar values with no outliers skewing the data or it graphs.

## Question 2

The data was prepared using the steps outlined in question one. Further preparation of the data was implemented by rescaling the variables and splitting them into either a test or train category. The data is split by putting one column into a variable named ‘X’, this is the independent column and the remaining data is put into a variable called ‘y’ which is the dependent column. The original variables are rescaled to give equal range or variance. The scale function performed on variable ‘X’ standardises the values within the column by using the following mathematical formula: X = X-mean(X)/std(X).

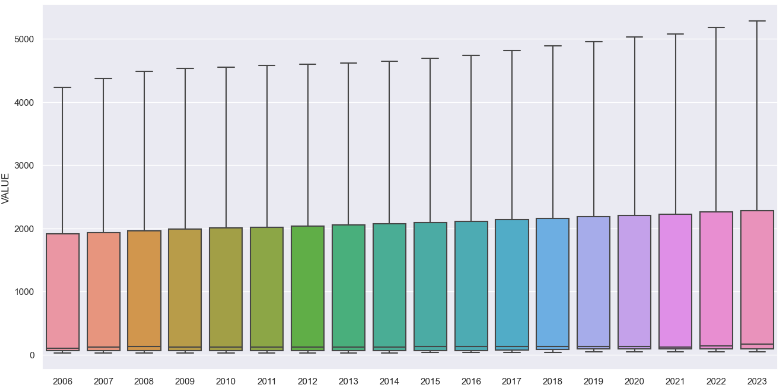
“Scaling your data in machine learning (ML)is important because many algorithms use the Euclidean distance between two data points in their computations/derivations, which is sensitive to the scale of the variables.”[2]

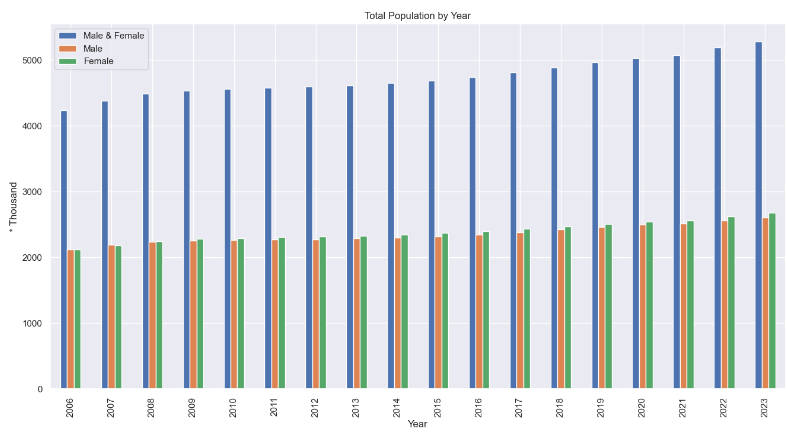
Now the X\_train, X\_test, y\_train, y\_test can be initialized by passing the X and y variables to the train\_test\_split function. The function will also take two other parameter which make four in total. The two other parameters are test\_size and random\_state. The test\_size parameter when set to 0.3 this tells then function that 30% of the dataset is to be allocated to the test set and the remaining 70% will be set to the training set. The random\_state parameter specifies how the data is to be shuffled as the data is passed to the variables. If set to ‘0’ the same results are obtained every time. In our case the random\_state was set to ‘4’. To finish check that the y\_train and y\_test variables have a similar distribution by printing out and comparing their mean value.

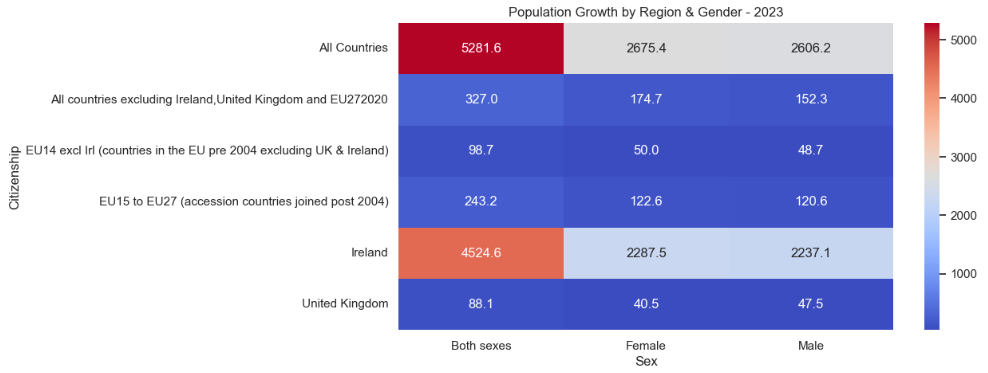
## Question 3

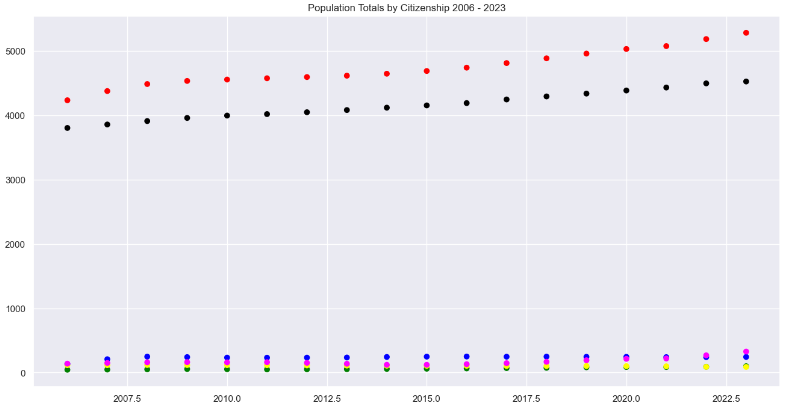
The boxplot was created by importing the Seaborn library. The boxplot itself is very basic and is used to visualize distributions. When comparing data between different groups the boxplot is quite useful as it does not overload the viewer and gives accurate information about the different groups within the dataset. The lines extending from the rectangles are called whiskers so, sometimes a boxplot may be called a box-and-whisker plot. The box/rectangle in the graph shows the quartiles of the dataset while the whiskers show the rest of the distribution. In the example below the data was split into each three groups with two of the groups representing genders and the third group represents both genders.

The vertical axis displays the population numbers and the horizontal axis displays the year those values represent. The lower whisker for each box shows the minimum value while the upper whisker shows the maximum value. Then the box itself is split between the first quartile and the third quartile with the median being displayed as a line separating the two quartiles. Each box represents eighteen rows of data.









# Statistics

## Question 1

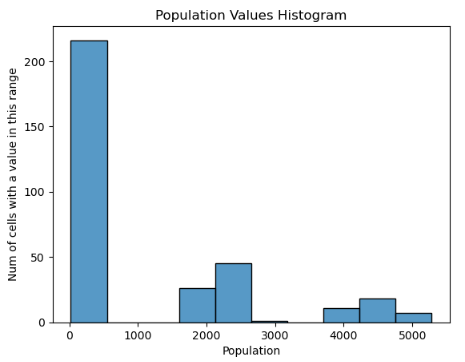
The dataset can be broken into two categories of data: The first category is categorical data which represent descriptive text used to give meaning to what is being viewed within the data set. The second category of data is numerical data. The numerical data itself can be of two types. The first type of numerical data is discrete data. Here the data can usually be counted in a finite matter while the second type of numerical data is continuous data. Here the data is infinite and impossible to count.

Using descriptive statistics, a data analyst has many techniques at the disposal which they may use to describe the data either numerically or graphically. Descriptive statistics usually involves using samples of dataset so the data itself can be transformed into information by organising the data. Having the data organised allows the data analysts to pull variables from the data such as averages that they may be used in functions provided by the various libraries available in python. Having the data organised also allows for the data to be used as an input for inferential statistics.

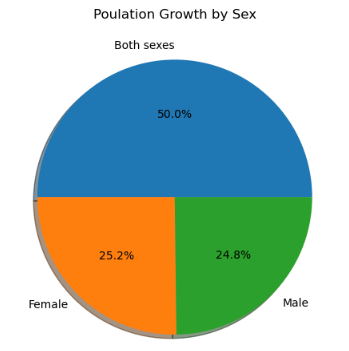
The data analyst needs to choose the right sample of data to obtain accurate information from the data set. Types of samples that may be used include a random sample, stratified random sample and cluster sample. An example of a random sample would be a weekly lottery draw. The result for that particular week would be random where as all the balls drawn out over the year would be a more complete sample. The stratified random sample is similar to the previous example except the data has been broken into groups before the random sample is taken. The cluster sample also breaks the data into groups but the difference here from the stratified random sample is there is individual samples are not taken instead groups of data are selected so all characteristic attributes of that data can be examined within the cluster.

Using these techniques, it was possible to create graphs and charts to give information about the data at a glance. First a histogram was created which shows the frequency of numerical data using rectangles. There were 324 rows of data within the spreadsheet with one column of numerical values. The histogram’s horizontal axis groups the numeric values column while its vertical axis is used to display the frequency of those values within the dataset. The histogram displayed below shows of the 324 rows of data over two hundred of those rows had a value less than one thousand (one million).

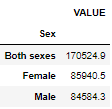
Note: For the dataset to be more readable, the dataset from the CSO displays its numeric values in units of a thousand. The values displayed need to be multiplied by one thousand to give their actual value. For example, 1000 population \* 1000 per unit = 1,000,000



The pie chart shows the totals for all the data within the dataset grouped by category. As the data for both sexes are combination of the male and female values it should equate to fifty percent while the values for the total amount of female and male data may differ form one another. Here we can see that there are more females in Ireland then there are males. From perform manual calculation this is the expected result. Also, from using Python to group the numeric values for each category of gender we are expecting to see the same result displayed within the pie chart.

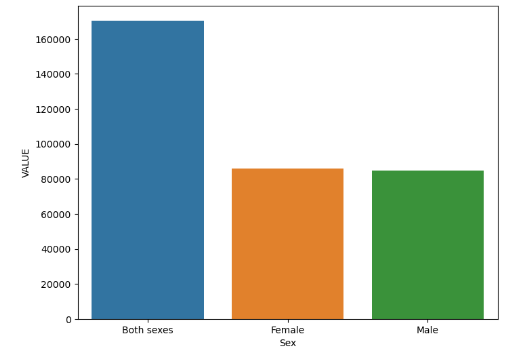


The bar chart’s vertical axis shows the total population values over eighteen years. The data within the CSO spreadsheet gives a total for each year and it also give the total numbers for each region. |With this in mind the figures could be halved as we collect the same data twice. Once as a total for all countries and a second time by summing the data from each region as a total. The horizontal axis shows the categories of data displayed within the chart. As the numbers for total females and males slightly differ it may be hard to see within the chart itself but there is a difference which can be seen in the below table followed by the chart itself.



Note: the values in the above table are double what the actual figure would be over the eighteen years. This is due to the data being collected twice. Once as a total for all counties and secondly as total from the following regions:

* EU14 excluding Ireland (countries in the EU pre 2004 excluding UK & Ireland)
* EU15 to EU27 (accession countries joined post 2004)
* United Kingdom
* Ireland
* All countries excluding Ireland, United Kingdom and EU272020



## Question 2

The Binomial Distribution is a discrete probability distribution. Distributions like Binomial tend to involve statistical analysis of counts to determine the probability of an event occurring. For this assignment Binomial distribution was used to find the probability of a given number of people being a certain gender within ten people. To get valid results for these probabilities three functions were used. These functions were made available after the binom library had been imported from scipy.stats.

The probability value used by the binomial functions was calculated by counting the total population and then dividing that result by the count for the total female population giving a number which represents the probability of the next person you meet in Ireland being a woman.

First the probability mass function was used. To use this function properly three parameters, have to be passed. The first parameter ‘k’ represents the amount of an occurrence we are looking for. The second parameter ‘n’ represents the total number of occurrences while the third parameter represents the probability of an event happening per iteration. A simple calculation to answer what the probability of three out of ten random people being female was.

The next function used was the cumulative distribution function. Using three parameters the probability of observing less than or equal to x can be obtained, ‘n’ represents the successes in trials, with the probability of success on a single trial being represented as ‘p’. To calculate the probability of having more than ‘k’ out of ‘n’ iterations the survival function was used which can be viewed as 1 – cdf.

The Poisson distribution also uses the same functions except here there are only two parameters passed. The first parameter ‘k’ represents the value we are looking to find the probability of by using the second parameter ‘mu’ which an average value pre calculated. To use Poisson first the average growth of the total population over eighteen years was found. This average was set as ‘mu’ and then ‘k’ was given a value of which we want to find the probability of. For example, ‘k’ was set to 100 which represents 100K population ‘mu’ the average growth value was ran against ‘k’ to give the probability to the total population within Ireland growing over one hundred thousand in a year.

## Question 3

Normal distribution also know as Gaussian distribution is a continuous distribution. Graphs created by using normal distribution tend to look like a bell and are commonly called “the Gauss Bell Curve”. The distribution takes into consideration an infinite number of data points where the expected value and variance are finite numbers. The distribution is symmetric so the curve displayed will never be to one side instead the expected average will be in the middle giving us the bell curve previously discussed. By importing the norm library from scipy.stats we gain access to the probability mass function, the cumulative distribution function and the survival function which were previously used for the Binomial and Poisson distributions.

## Question 4

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case.

# Programming

## Question 1

For this assignment the Python programming language was used to import data, manipulate the structure of the data, perform calculations on the data and then present the results from the calculations graphically or numerically. IPython stands for interactive python and provides a terminal with a web-based notebook for computing python. In the python files it was directly used to import the Image library so images could be displayed within the Juypter notebook as well as being the backend for python itself.

The Python programming language was created in 1991 by Guido Van Rossum at CWI Netherlands. Python itself is a high-level programming language that is general-purpose and somewhat easy to learn as it is concise and easy to read. The language does not depend on a specific platform to run and is free to use as well as being open source which allows for the community that maintain the language to add improvements to the language as needed. The code is written as a list of commands into scripts with the .py file extension and will run from the first line/command to the last.

“In 2015, the Ipython developers made a significant code reorganization of their project. So, the Notebook is now called the Jupyter Notebook. So, this interface is used with Python and many languages like R and Julia. IPyhton is the name of the Python backend.”[1]

In order to complete the assignment, the following libraries were used:

Pandas is an open-source library and allows for the data structure to be manipulated for quick analysis and easy viewing. The Numpy library was used to store the data in multidimensional arrays and matrices. Numpy provides high-level mathematical functions which allow for calculations to be performed on the stored data. The Seaborn library was used for visualizing the results of the data in many formats such as heat-maps, histograms, bar charts and boxplots. Seaborn is derived from Matplotlib and is integrated with the Pandas data structures. The Matplotlib library was used for dimensional plotting, this means results can be viewed interactively as opposed to viewing results statically from a notebook. The statistics library when imported offered a range of functions for calculating the statistical data. Using the Statistics library with the Scipy library allowed for the importing of the binom, poisson and norm libraries which where needed to complete questions two and three of statistics section of this assignment.

“Scipy is also used for Data Computation, productivity, high-performance computing, and quality assurance. The various installation packages can be found here. The core Scipy packages are Numpy, SciPy library, Matplotlib, IPython, Sympy, and Pandas.” [2]

Scikit- learn

It is a free software machine learning library for the Python programming language. It can be effectively used for a variety of applications which include classification, regression, clustering, model selection, naive Bayes’, grade boosting, K-means, and preprocessing.

## Question 2

No specific programming paradigm was chosen when beginning this assignment. However, after the programming aspects of the assignment were completed an imperative programming paradigm was visible due to each instruction being part of a set instructions to yield the desired result when executed in a sequential order. Writing the code like this solved the problems asked in the assignment as the computer was told exactly what to do and when to do it.

Other options of programming paradigms would include but not limited too are functional, and object oriented. Had these paradigms been used it is possible the code would be cleaner and easier to read as the repetition of repeating cells may have been removed due to the code only having to be wrote once in a secondary file and them imported to the main program file for use. Also object orientated programming uses inheritance which hides the data types and keeps the data secure which is desirable.

“Programming paradigms are different ways or styles in which a given program or programming language can be organized. Each paradigm consists of certain structures, features, and opinions about how common programming problems should be tackled.”[3]

Quaid, D.M. (2023) ‘Lecture 2 - What is Exploratory Data Analysis?’, *Data Preparation and Visualisation*. *Online Class*, Dublin: CCT, 26 September.

Khoong, W.H. (2023) *Why scaling your data is important*, *Medium*. Available at: https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2#:~:text=Scaling%20your%20data%20in%20machine,the%20scale%20of%20the%20variables. (Accessed: 16 November 2023).

*IPython display - javatpoint* (2021) *www.javatpoint.com*. Available at: https://www.javatpoint.com/ipython-display#:~:text=It%20is%20an%20interactive%20command,single%20line%20of%20Python%20code. (Accessed: 16 November 2023).

Team, G.L. (2023) *Top 30 python libraries to know in 2024*, *Great Learning Blog: Free Resources what Matters to shape your Career!* Available at: https://www.mygreatlearning.com/blog/open-source-python-libraries/#:~:text=Python%20Libraries%20are%20a%20set,data%20manipulation%20applications%2C%20and%20more. (Accessed: 16 November 2023).

Cocca, G. (2022) *Programming paradigms – paradigm examples for beginners*, *freeCodeCamp.org*. Available at: https://www.freecodecamp.org/news/an-introduction-to-programming-paradigms/#:~:text=What%20is%20a%20Programming%20Paradigm%3F,programming%20problems%20should%20be%20tackled. (Accessed: 16 November 2023).