Project Report

GitHub URL

<https://github.com/patriciawilson2021/UCDPA_Patricia-Wilson>

Abstract

(Short overview of the entire project and features)

My project is an analysis of Medal winners during the Summer Olympic Games from 1896 – 2012. It incorporates 6 different types of graphs as well as using a table scraped from the web to show medals won by Ireland in the 2020 Olympic Games.

The project is comprised of the following sections, and analyses the data to answer the questions marked 1-6 below.

Data Import  
Data Preparation

1. Do countries with greater GDP win more medals?
2. Count Gold, silver, Bronze medals by Gender
3. Who are the top 10 most successful athletes at the Summer Olympics from 1896 -2012?
4. How many Gold, Silver and Bronze Medals have been won by Ireland between 1896 - 2012?
5. In which years did Irish Olympians win medals?
6. How many medals were won by Ireland in the 2020 Tokyo Olympics?

Define a custom function to get min, max and average populations from the complete country details DataFrame and print the results

Introduction

(Explain why you chose this project use case)

I chose to use this project use case as I found the subject matter interesting and I could see that the underlying datasets provided scope for using different types of graphs and charts in the analysis.

Dataset

(Provide a description of your dataset and source. Also justify why you chose this source)

The dataset I used came from https://www.kaggle.com/the-guardian/olympic-games. Kaggle is a popular data repository and the dataset had a usability score of 7.1 which meant that the data:

* Is easy to understand and includes essential metadata
* Has rich machine readable file formats and metadata
* Assures the dataset is maintained

The dataset consists of 2 files. Originally these were .csv files but I had to convert them to an .xlsx file type to successfully import them into pycharm. The files are stored in my github repository from where they are imported into pycharm.

Dataset 1: ‘Summer.xlsx’

This file has 9 columns and 31165 rows. It contains details about each Olympic Games, and the Athletes who won medals in each event in the games. The breakdown of the columns is as follows:

* **Year** – The year in which the Olympic Games was held
* **City** – The City in which the Olympic Games was held
* **Sport** – The sport being participated in
* **Discipline** – the subset of the sport being participated in
* **Athlete** – The name of the athlete participating
* **Code** – a field consisting of a 3-letter description of the country the participating athlete is representing
* **Gender** - shows ‘Men’ or ‘Women’ depending on the gender of the event
* **Event** – the event being participated in. This is a further subset of Sport and Discipline.
* **Medal** – shows whether a Gold, Silver or Bronze medal was won in that event by the participating athlete

Dataset 2: ‘Dictionary.xlsx’

This file has 4 columns and 201 rows. It contains details about the Countries participating in the Olympic Games. The breakdown of the columns is as follows:

* **Country** – The name of the country
* **Code** – a 3 letter description of the country
* **Population** – the population of the country
* **GDP** – the GDP per capita of the country

Dataset 3: Tokyo Olympics Ireland Medal Table

This table was scraped from Wikipedia using beautifulsoup. I used Wikipedia as the source of this table because as the information shown in the table on the site was accurate and it provided me with the opportunity to use beautifulsoup to scrape the table from the website.

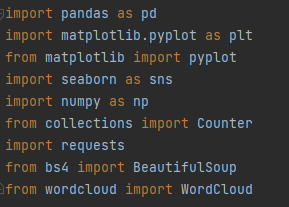
It contains 5 columns and 2 rows. The columns are as follows:

* **Sport** – Name of the sport in which medal was won by an Irish athlete
* **Gold** – How many gold medals were won
* **Silver** – How many Silver medals were won
* **Bronze** – How many Bronze medals were won
* **Total** – total number of medals won in this sport

Implementation Process

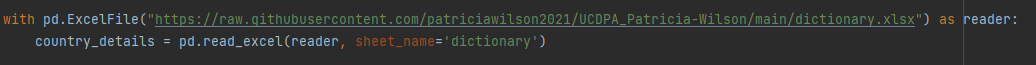
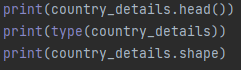
(Describe your entire process in detail)

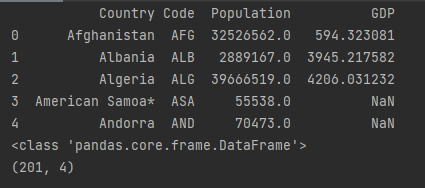
Imported modules



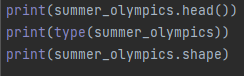
Data Import

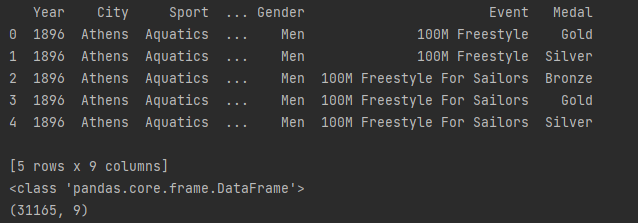
Dictionary.xlsx and summer.xlsx are imported into pycharm and converted to DataFrames.

Load Dictionary.xlsx and convert to DataFrame (country\_details). A DataFrame is the most suitable structure to load the data into for analysis to be carried out.  
  
  
  
Show details of country\_details DataFrame  




* Head of country\_details dataframe
* Type: Pandas Dataframe
* 201 rows, 4 columns in Dataframe

Load summer.xlsx. Convert to DataFrame (summer)  
  
View head of DataFrame  


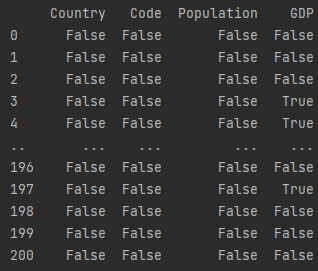


* Head of summer dataframe
* Type: Pandas Dataframe
* 31165 rows, 9 columns in Dataframe

Data Preparation

Check country\_details DataFrame for empty values  

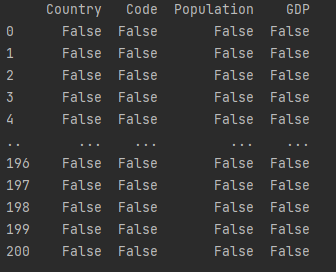

The country\_details Dataframe contains null values in the GDP column

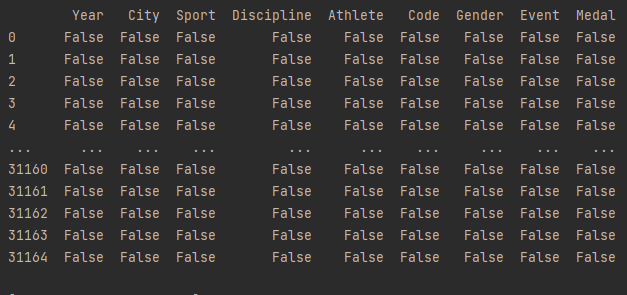


Use fillna(0) function to replace nan values. Missing values are replaced with 0 to ensure a complete dataset is used in the DataFrame.  


Check complete country\_details to make sure that no nulls remain  


No Nulls found in country\_details DataFrame

  
  
Check summer\_olympics for null values. None found.  

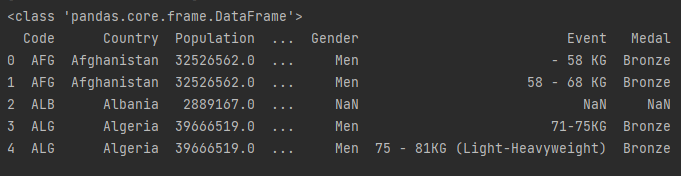



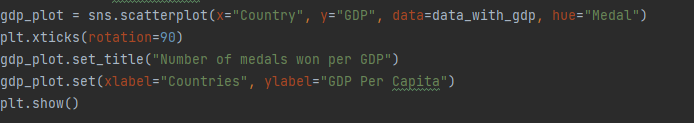
1. Do Countries with greater GDP win more medals?

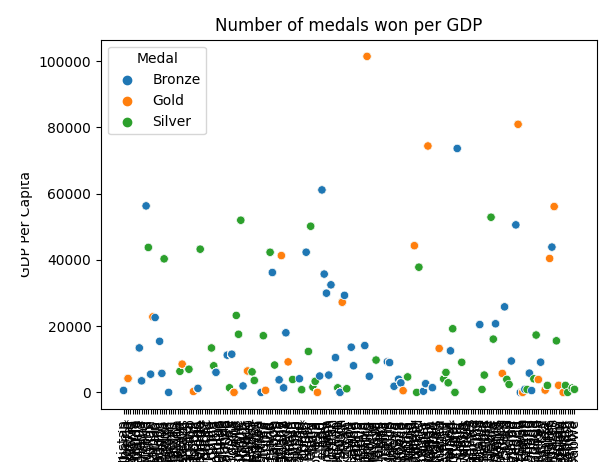
Set index of country\_details to 'code'.  
  
  
Merge country\_details and summer\_olympics dataframe into new dataframe called 'data\_with\_gdp' to allow information to be extracted from both dataframes for visualisation. A left join is used to extract all rows from country\_details\_ind and matching rows from summer\_olympics. Code is used as the primary key as it is a unique identifier in both DataFrames.  

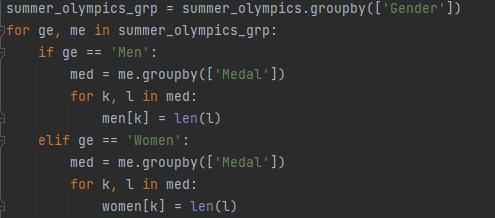

Print the type and head of data\_with\_gdp

Print the head of the DataFrame  

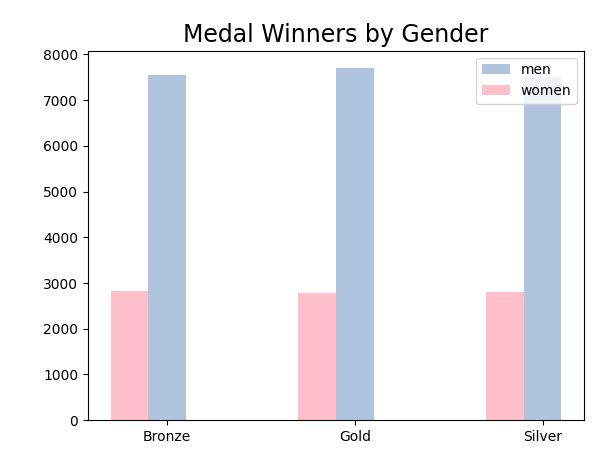
Use Seaborn scatterplot to visualise data.  




2. Count Gold, silver, Bronze medals by Gender  
  
Create separate dictionaries for medals won by men and women. Dictionaries were used as I wanted to store key-value pairs.  
  
  
Loop through summer\_olympics dataframe to populate dictionaries and count values of gold, silver & bronze medals. Groupby is used to split data first into Gender and then Medals to create 2 dictionaries of medals won by each gender  


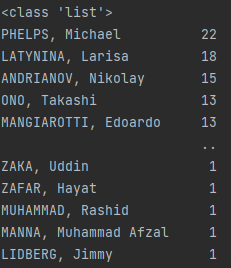
Print Dictionaries for men and women  


  
  
Plot information on grouped bar chart.  

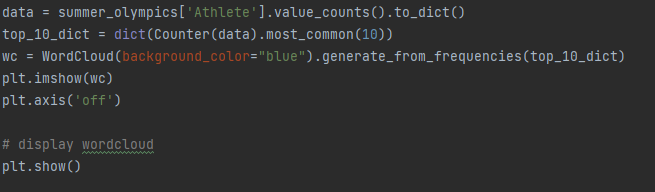
3. Who are the top 10 most successful athletes at the summer olympics from 1896 -2012?

Extract athletes column from summer\_olympics dataframe. As only one element is to be extracted, a list is used.  

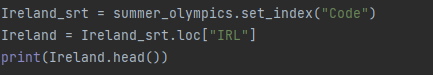

  
  
Count how many times each athlete appears in the list  

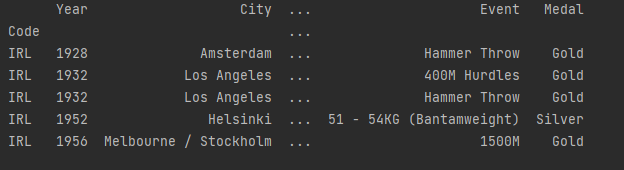

Create list of the top 10 rows from the count\_top10 list.  


  
  
Create a word cloud visualisation. I used a Datacamp tutorial on Generating wordclouds in Python to discover how to create this type of visualisation. https://www.datacamp.com/community/tutorials/wordcloud-python

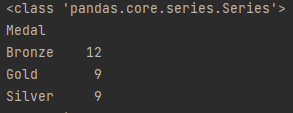
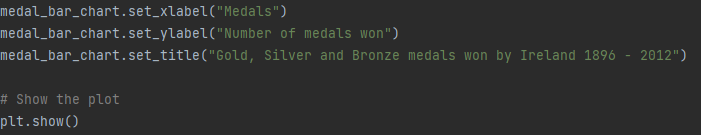


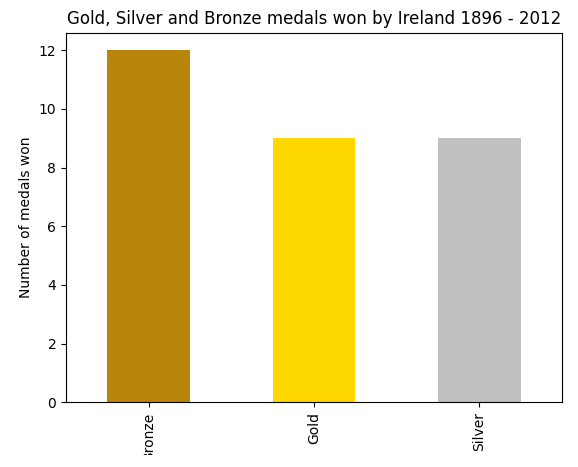


4. How many Gold, Silver and Bronze Medals have been won by Ireland between 1896 - 2012?  
  
Set index of DataFrame to ‘Code’. Use .loc to extract all rows where code is 'IRL'. .loc is used as it is a method in Pandas to extract rows where the indexed label exists.  


  
  
Sum of each medal type won.   

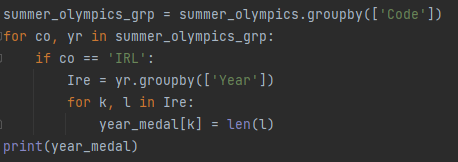

Print the type of medal\_totals, Returns a series.  


  
  
Create bar plot of number of medals won by type  


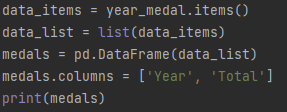


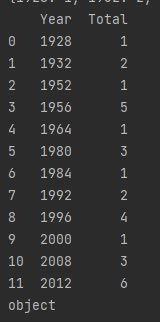
5. In which years did Irish Olympians win medals?  
  
Create dictionary of years where medals were won by Ireland. I used a dictionary here to store both the years where Ireland won medals and the number of medals won in each of those years. As this is a key value pair, a dictionary is the most suitable storage method.  

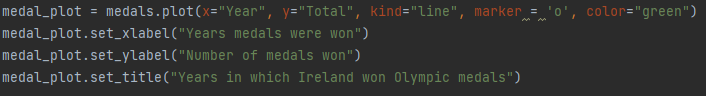

Group summer\_olympics DataFrame by ‘Code’  


Loop through grouped DataFrame to extract years & Medals won. Groupby is used to split data where code is Ireland, and then the number of medals in each year  




Convert dictionary to DataFrame for visualisation. The DataFrame is called medals and it has 2 columns called ‘Year’ and ‘Total’, to hold the years and the number of medals in each year respectively.  
  
  
Check type of items in list  

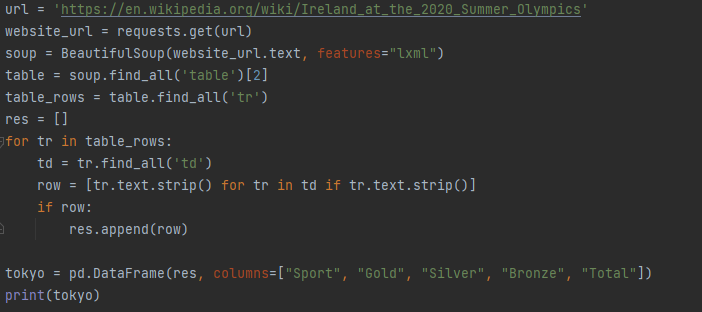
Convert year column to numeric to allow it to be used as x axis  
  
  
Create medal\_plot  
  
  
Set x axis to show all years (in steps of 4) from first year in which Ireland won medals to the last.  

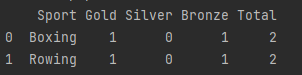

Add grid to make it easier to match x and y values  


Show Plot  

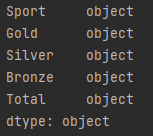


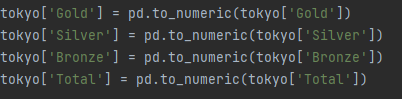

6. How many medals were won by Ireland in the 2020 Tokyo Olympics?

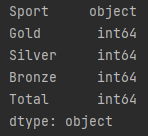
Use beautifulsoup to scrape a table from Wikipedia. This library is specifically created for pulling data from webpages so was the most suitable method to use*.*  


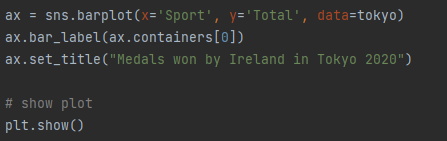


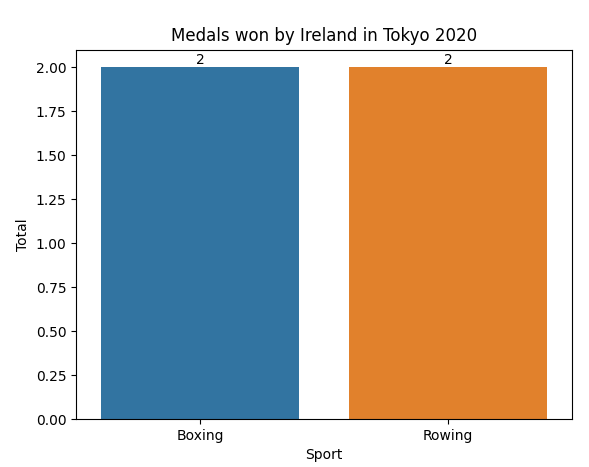
Check data types of columns  


  
  
Columns are displaying as objects so need to change these to numerical values for visualisation

  
  
Print types of columns again to ensure type has changed  

Create visualisation of number of medals won in each sport by Ireland in Tokyo 2020  


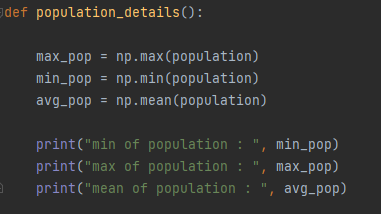


Define a custom function to get min, max and average populations from the complete country details dataframe and print the results

The function is defined as ‘population\_details’.

It returns the max, min and average of a set of population details and populates the variables ‘max\_pop’, ‘min\_pop’, ‘avg\_pop’ with these figures.

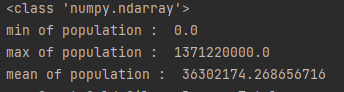
Finally, the function prints these details to screen.



Create a NumPy array of the population column from the complete country details DataFrame. I used a NumPy array as the mean, min and max aggregations are part of the NumPy function.

  
  
Call the population\_details function to carry out calculations on the population array  


Results:



Population\_details is a numpy.ndarray

The country with the smallest population was 0

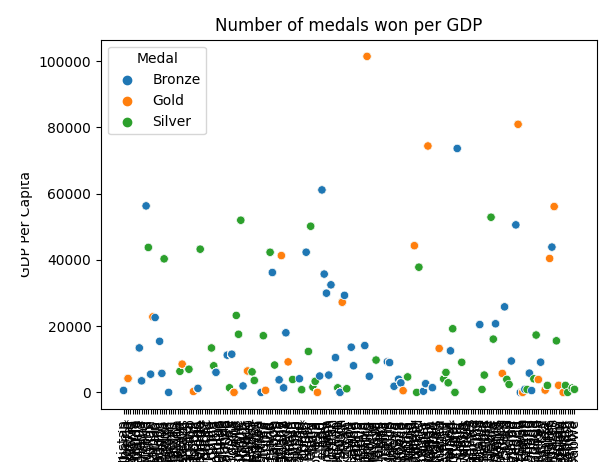
The largest population was 1371220000.0

The average population was 36302174.268656716

Results

(Include the charts and describe them)

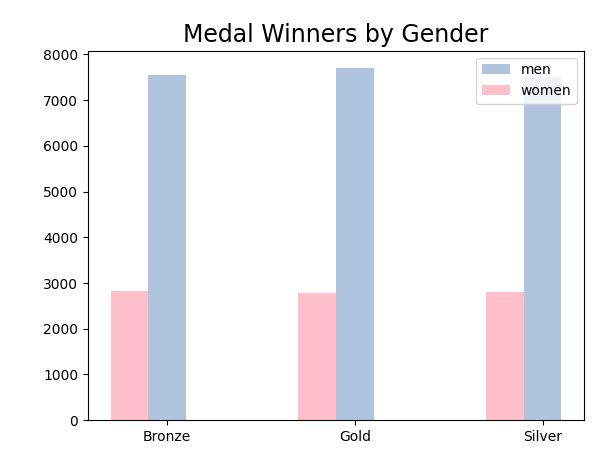
1.



A Seaborn scatterplot is used to show relationships between two numeric variables, for example number of medals and GDP per capita.

The X axis shows the country names. The X axis shows the GDP per capita. The medal type was used for the hue to show the distribution of medal type per capita.

2



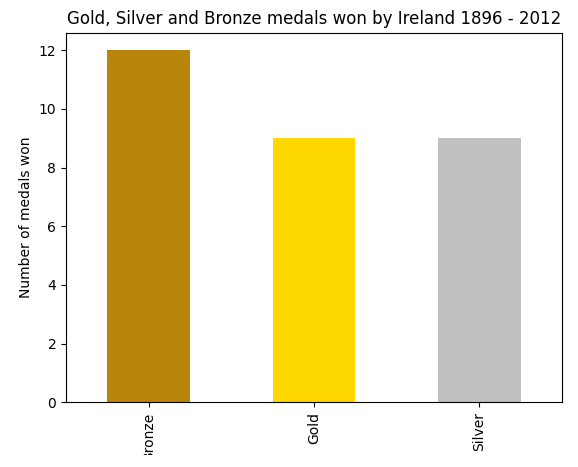
A grouped bar chart was used in this example to provide a comparison in the performance of Male and Female athletes in each medal category. The x axis shows Gold, silver and bronze medals for both male and female athletes. The legend shows pink columns for women and blue for men. The Y axis shows the number of medals won in each category.

3



A wordcloud provides an engaging method of displaying the data visually. The text size reflects its frequency in the dataset used, where the text in larger font occurs more frequently. Wordclouds were not covered in the course but researching this visualisation type allowed me to expand upon my learning in the course.

4



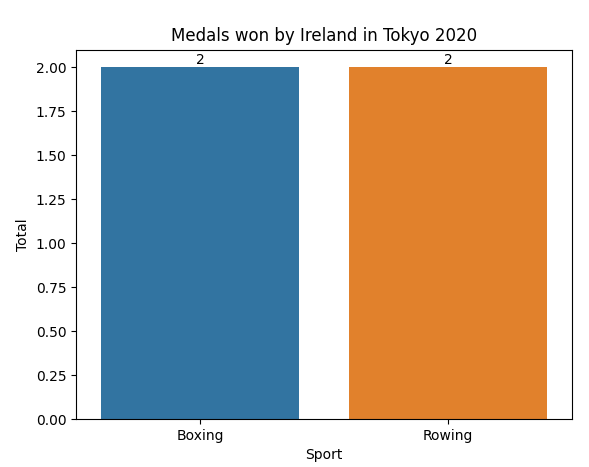
This matplotlib bar chart shows that Ireland have won a total of 12 bronze medals as well as 9 gold and 9 silver medals between 1896 – 2012. Using a bar chart makes it easy to compare the numbers in each medal category.

5



A line chart shows changes in data over time. The X axis shows the years of each Olympics from the first year medals were won by Ireland. The Y axis shows the number of medals won. The markers represent the number of medals won by Ireland in that particular year and a grid was added to make the correlation between the two variables easier to identify.

6



A Seaborn barplot was used to represent the data in this visualisation to display the totals at the top of each bar. The Sports in which medals were won are displayed on the X axis. The number of medals won are shown on the Y axis.

Insights

(Point out at least 5 insights in bullet points)

* The higher the GDP per capita, less medals are won
* Men won significantly more medals than women in each medal category
* Michael Phelps is the most successful Olympian in Olympic history up to 2012
* Between 1896 – 2012, Ireland have won 12 bronze, 9 gold and 9 silver medals
* Ireland’s most successful Olympic year was in London in 2012, where 6 medals were won
* In Tokyo 2020, Ireland won 4 medals, 2 in Boxing and 2 in Rowing.

References

(Include any references if required)

Vu, Duong, November 2019, Generating Wordclouds in Python, Datacamp, viewed November 2021, <https://www.datacamp.com/community/tutorials/wordcloud-python>

Hunter, John, Dale, Darren, Firing, Eric, Droettboom, Michael, List of Named Colours, Matplotlib, 2021, viewed November 2021, https://matplotlib.org/stable/gallery/color/named\_colors.html