

IST652 Scripting for Data Analysis

**Sport Injustice – Olympics History as an example**

**Data Analysis on Olympics dataset and Countries Economic Metrics**

**Final project**

**December 2022**

**SUBMITTED BY** Somia Abdelrahman

**Introduction:**

The modern Olympic Games or Olympics are the leading international sporting events featuring summer and winter sports competitions in which thousands of athletes from around the world participate in a variety of competitions. The Olympic Games are considered the world's foremost sports competition with more than 200 teams, representing sovereign states and territories.

In our final project we plan to perform exploratory Data Analysis using Python to analyze and visualize past Olympics data and gain valuable insights from it. We also want to explore trends in sport and examine the complex connections between sport, economics and gender. This analysis is expected to help us identify biases and inequality in sport through Olympics history.

**Datasets:**

We used two datasets in this project. The first one was a structured csv file which contained the data of 120 years of Olympic games history. This csv file had a lot of information including the name of the participant, his age, number of medals won, nationality, sport that he played, gender, height, weight and more.

Glimpse of the information:



This data and the columns such as Team, NOC, sport, Medals won, Sport, Games will be very useful to perform analysis for our questions and find the correlation between the Olympics and the role of economics of a particular country.

The link for this dataset is:

<https://www.mavenanalytics.io/data-playground?page=3>

The second dataset that we used is a semi-structured data set which contains information like Country’s demographic and economic metrics such as Population, Gross Domestic Product. We fetched this data from the world bank website using their API.

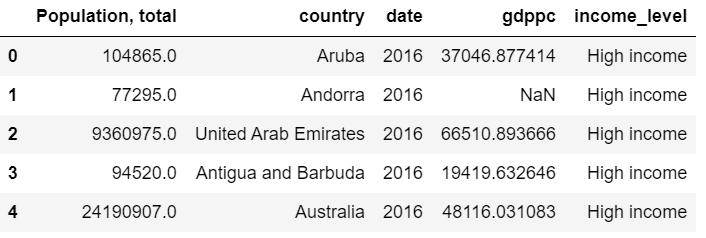
Now let’s talk exactly what are Country’s Demographic and about Gross Domestic Product (GDP):

Demographics are statistics that describe populations and their characteristics. Demographic analysis is the study of a population-based on factors such as age, race, and sex. Demographic data refers to socioeconomic information expressed statistically, including employment, education, income, marriage rates, birth and death rates, and more.

From this information it’ll be very important to understand the income levels of different countries along with their population and GDP.

Well here Gross Domestic Product plays a crucial role as it is an important indicator of economic performance and a useful unit to make cross-country comparisons of average living standards and economic wellbeing.

Glimpse of the Data frame:



The link from where we retrieved the data is:

<https://data.worldbank.org/>

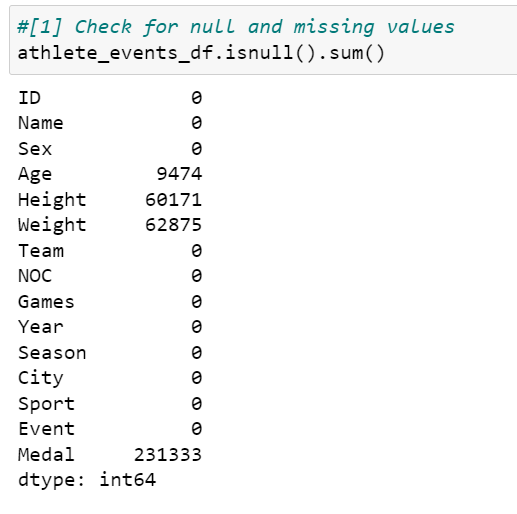
So after skimming through these 2 datasets we decided to perform an analysis and find insights for topics like gender distribution, maximum medals won (female participants), economic influence on Olympic games and the medals won and more.

So before jumping on to the Analysis, it was important to clean the data and drop un-necessary columns as they weren’t needed.

**Data Exploration:**

For 120 years of Olympic games history, there was a need to load 2 datasets because one dataset contained all the athlete’s information and the other one contained the country definitions. So merging these 2 datasets in a single data frame.

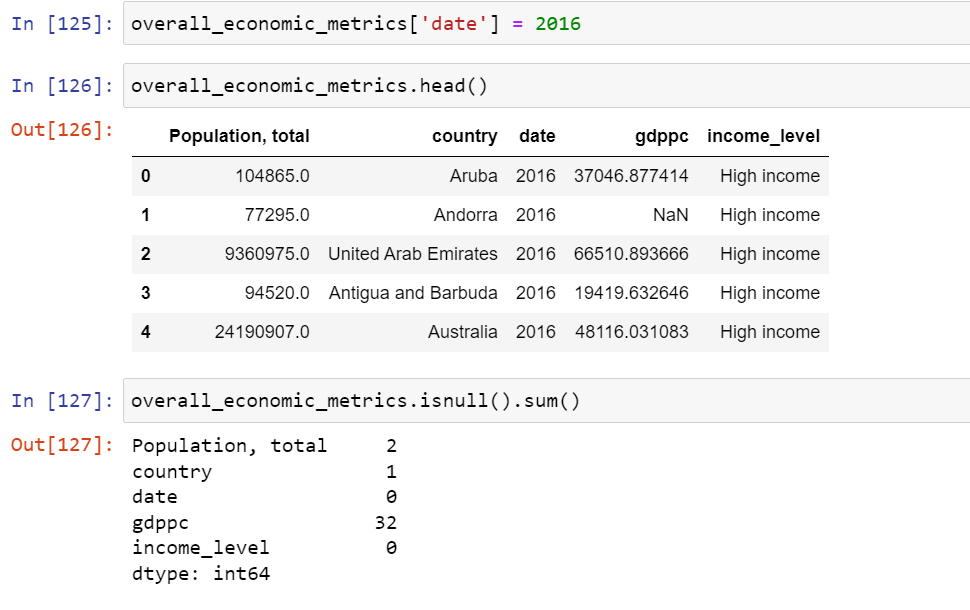
The first thing that we need to do is to check for NAN values and the data types of the information in every column that we will use for analysis.

 So we can see that there are total 9474 missing values in Age, 60171 in height and 62875. We can drop these NAN values. But it’s very important in an analysis to understand which columns we are going to use and in our analysis part we are going to focus only on the columns such as Team, NOC, Games, Season, City, Sport, Event and Medal.



So instead of dropping Nan Values just updated my data frame with the columns needed. There was also a need to fetch information for any particular year’s information for Olympic games so that we could find a correlation between the Olympic games and the economics metrics for that particular year.

So as discussed before in the datasets part we fetched the data from a website using their API which contained information like Country’s demographic and economic metrics such as Population, Gross Domestic Product.



After retrieving the data we just search for the Null values and drop these values and start with our Analysis.

Moving on to the Analysis part.

# **Description of Methods of Analysis:**

This project intends to examine the complex relationship between sport and gender by evaluating female participation in sport over the years and how the economic status of certain country can affect its opportunity to make achievement in the international sport tournaments.

Hence, the aim of this project is to answer below questions:

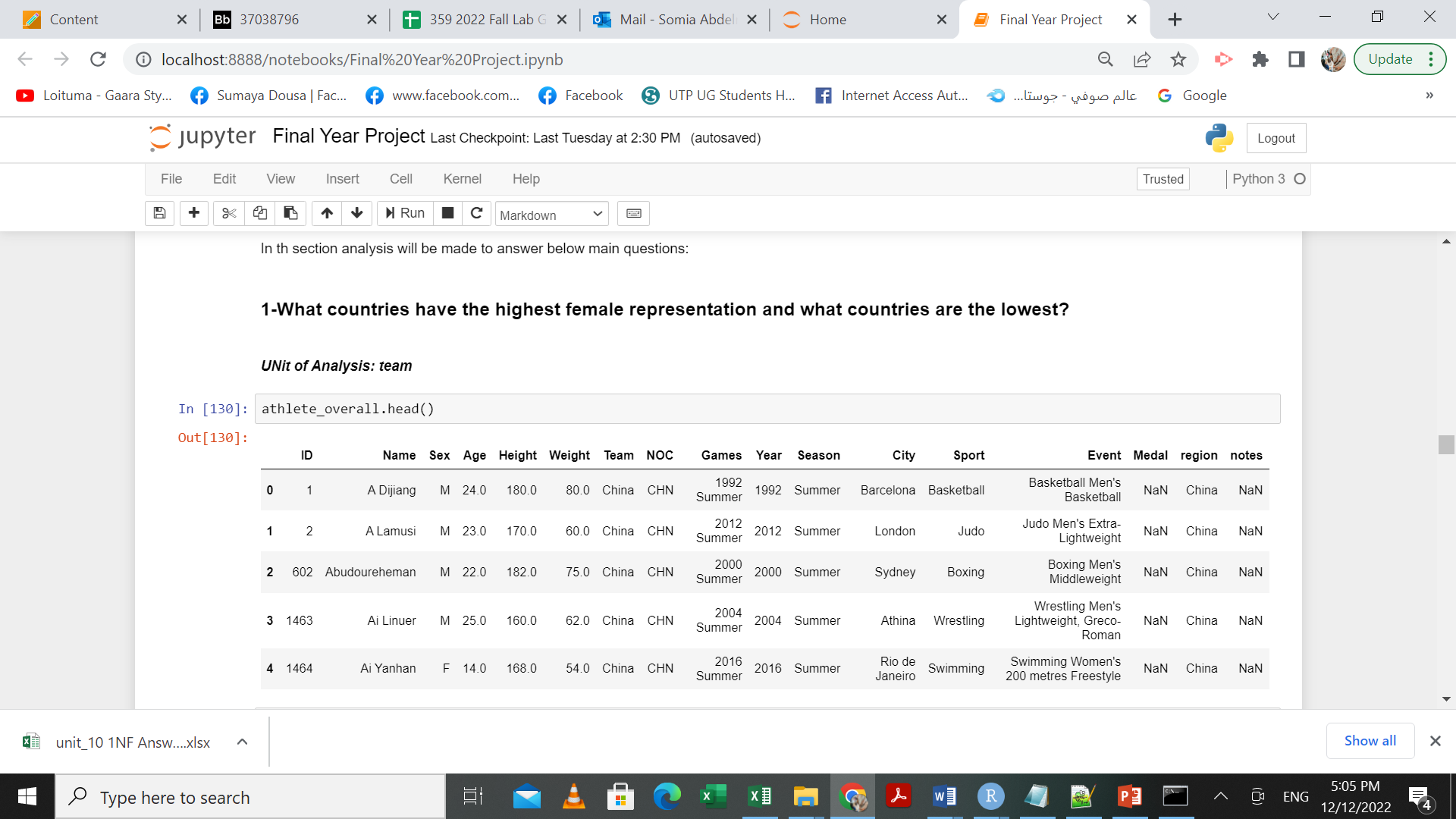
* What countries have the highest female representation and what countries are the lowest?
* What is the recent percentage of women athletes in the Olympics and how did it change over time?
* What are the top five countries in terms of No of medals achieved throughout the Olympics history and what is the percentage of female representation in these teams?
* Does country GDP and expenditure on sport influence the number of medals achieved? Is there any correlation between spending on sport and the number of medals?

In below paragraphs, we will discuss in details each question and provide the results we achieved throughout the analysis.

## **What countries have the highest female representation and what countries are the lowest?**

Unit: Female athlete per country

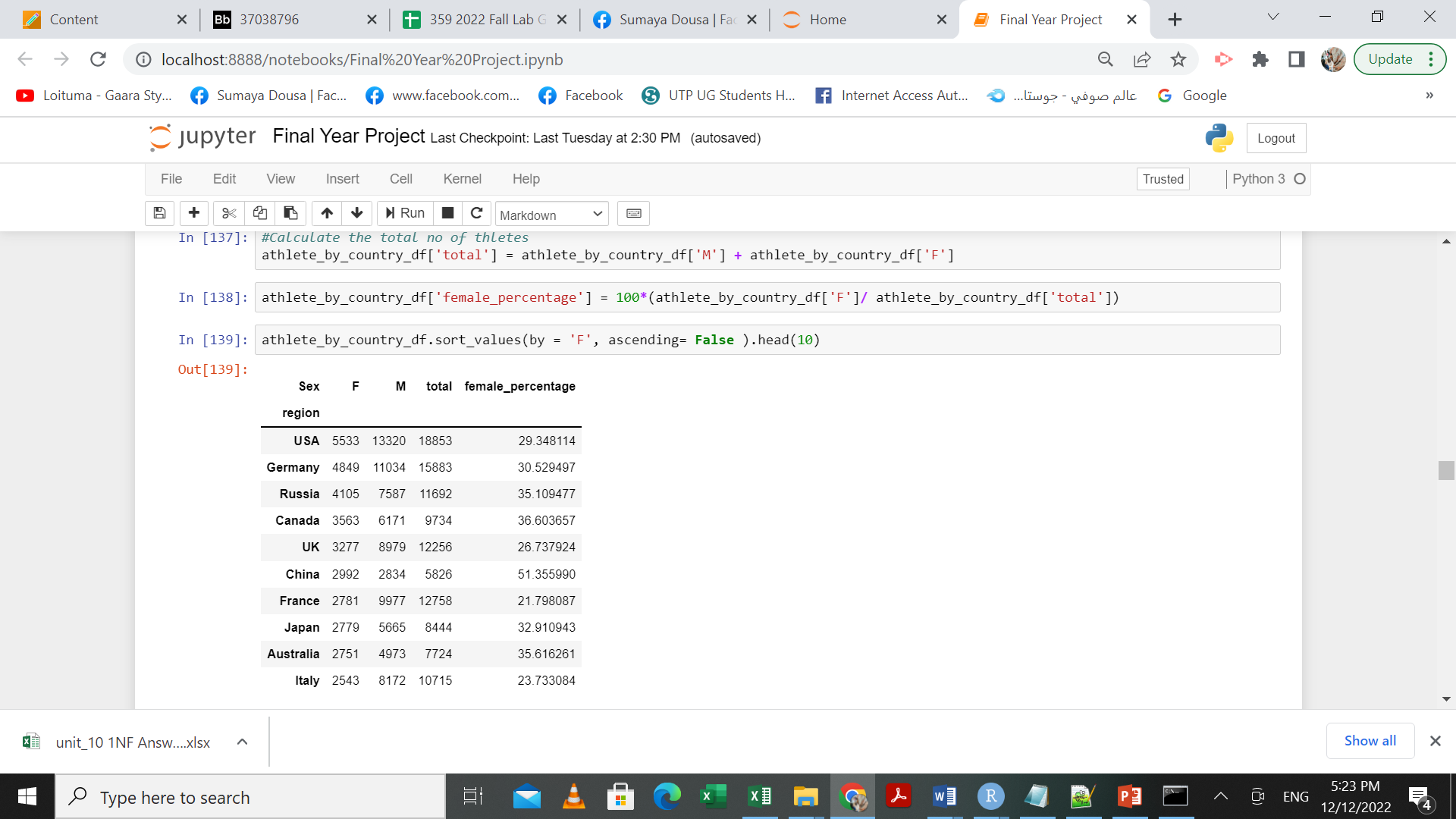
To answer this question, the Region and the Sex fields were used. The region column represents the country that participant belongs to.

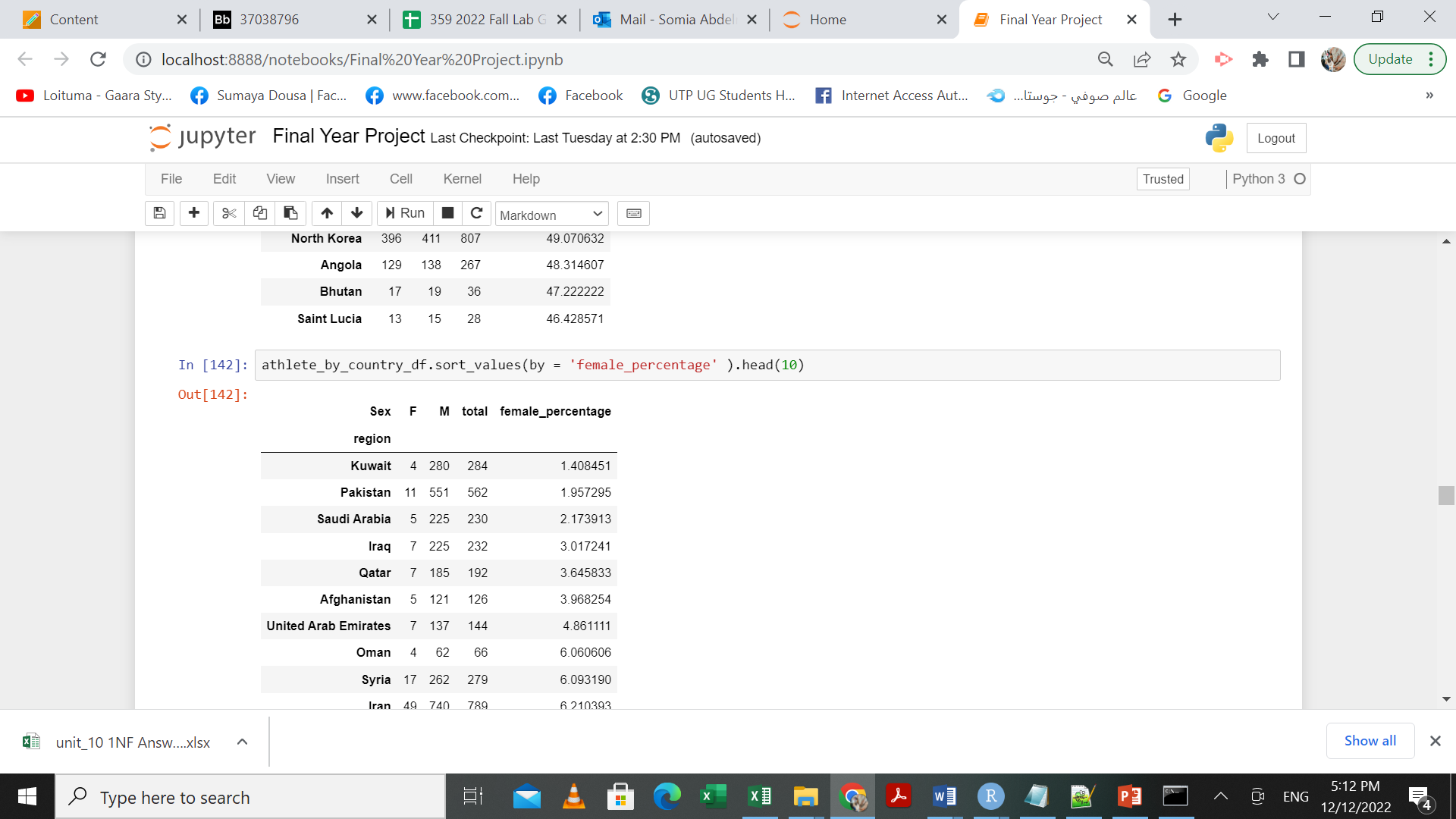


Next, we aggregated the dataframe based on the column of the region and then calculated the number of Female and Male in these countries using the value\_counts() function. Finally we calculated the female participation percentage in each country based on the following:

Female Percentage = Female total participants/ Overall participants

And the final results were as follows:

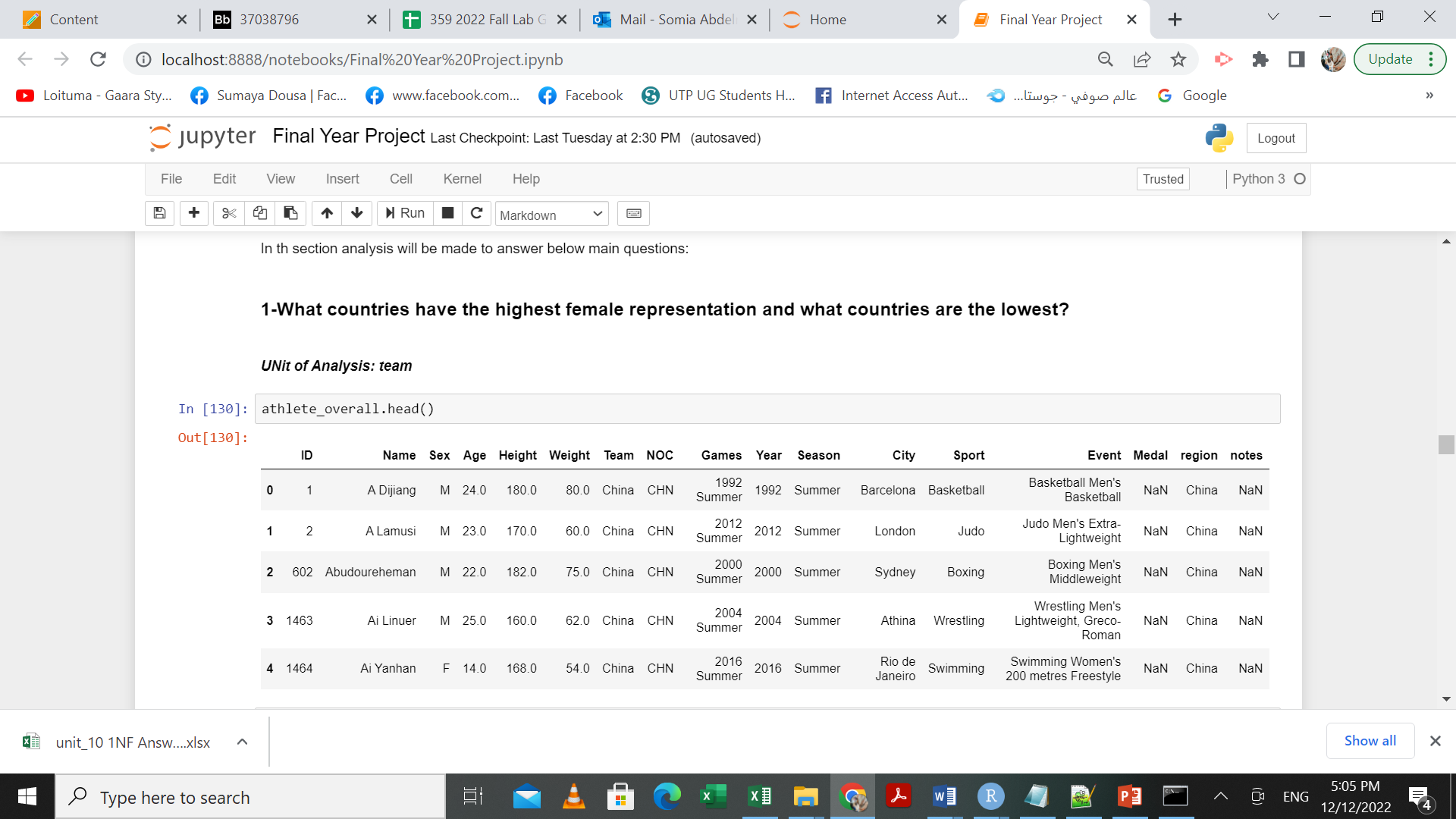




## **What is the recent percentage of women athletes in the Olympics and how did it change over time?**

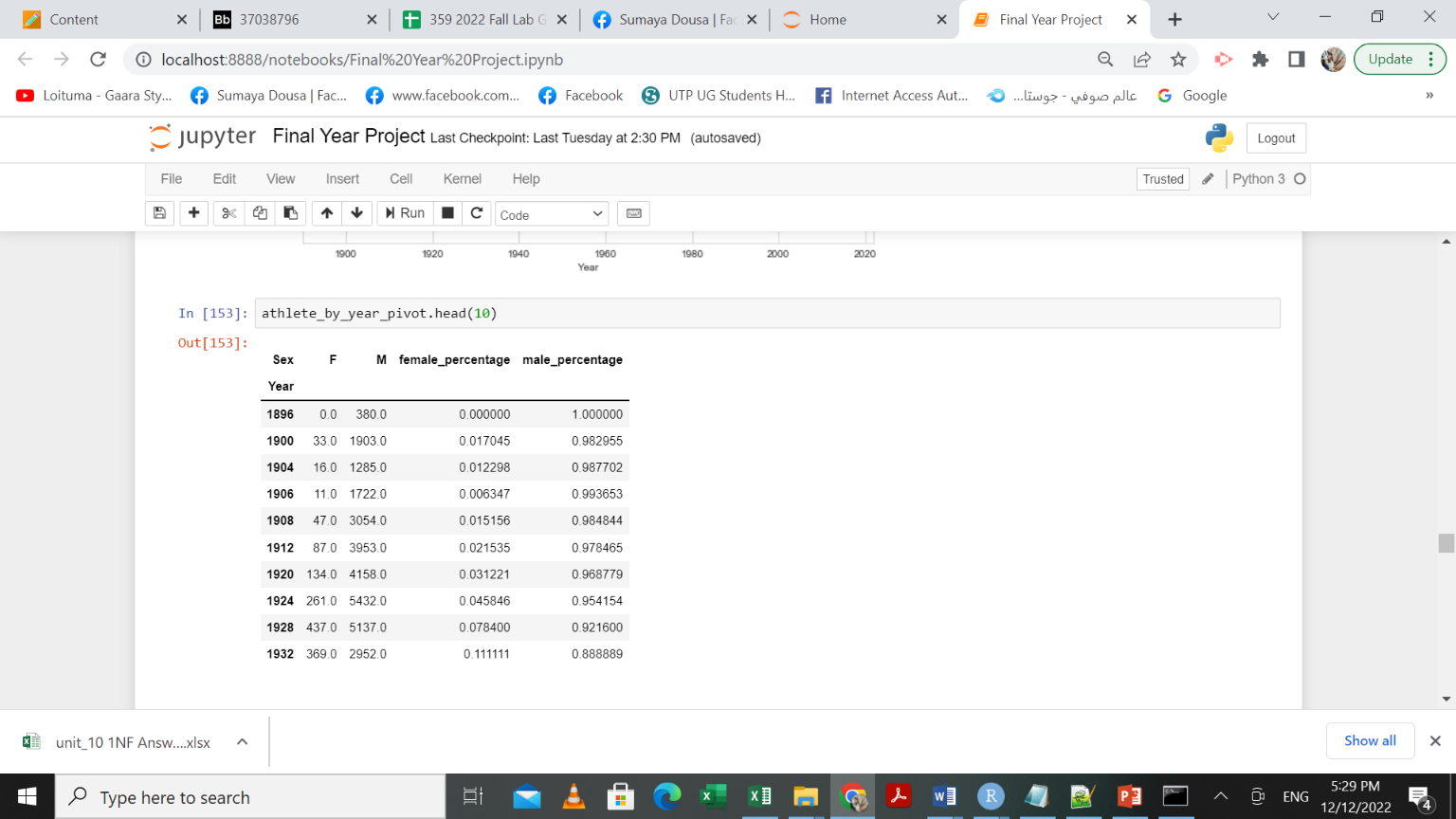
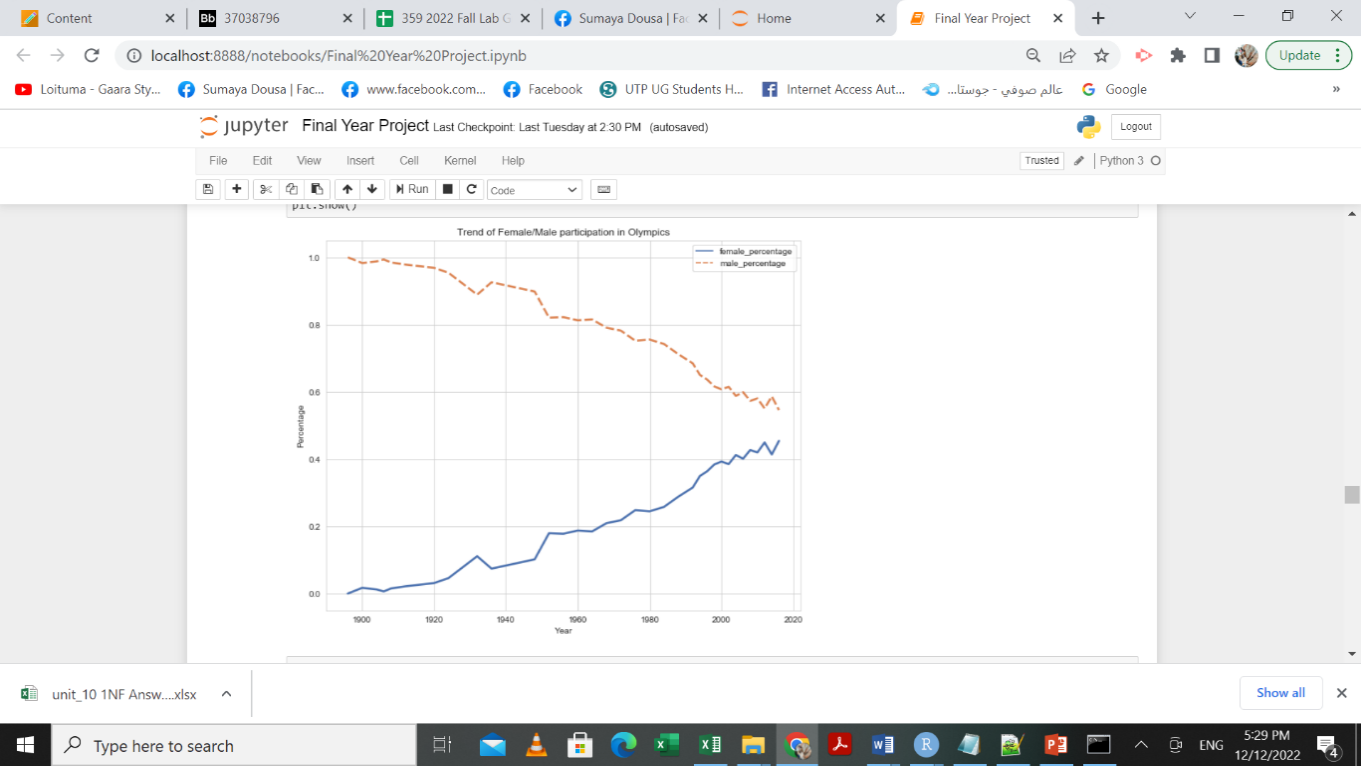
Unit(s): Percentage/year

To answer this question, the Year and the Sex fields were used. The year column represents the year that athlete participated in.



Next, we aggregated the dataframe based on the column of the year and then calculated the number of Female and Male in who participated in these years using the value\_counts() function. Finally we calculated the female participation percentage in each year.

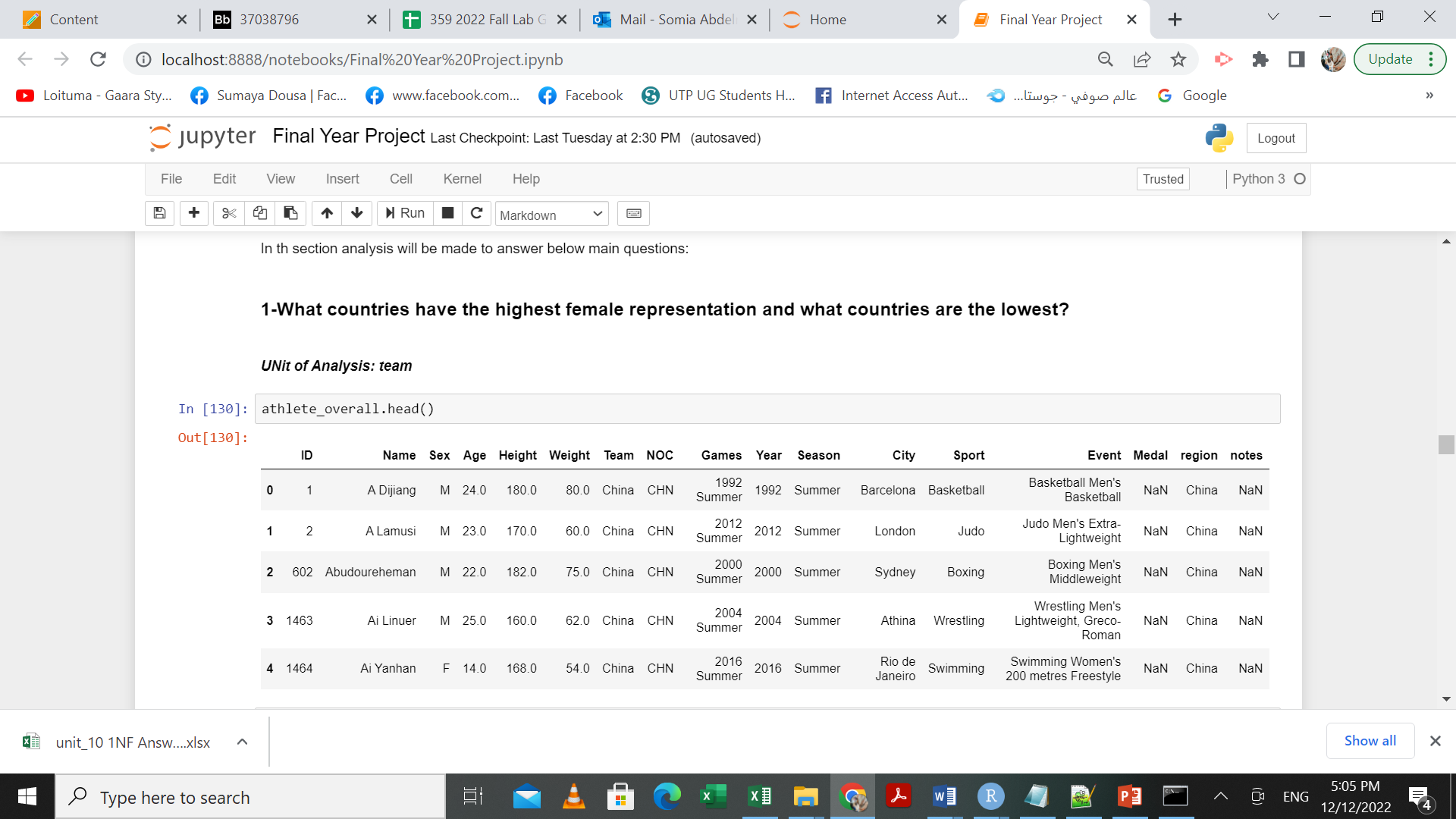
The results were as below:



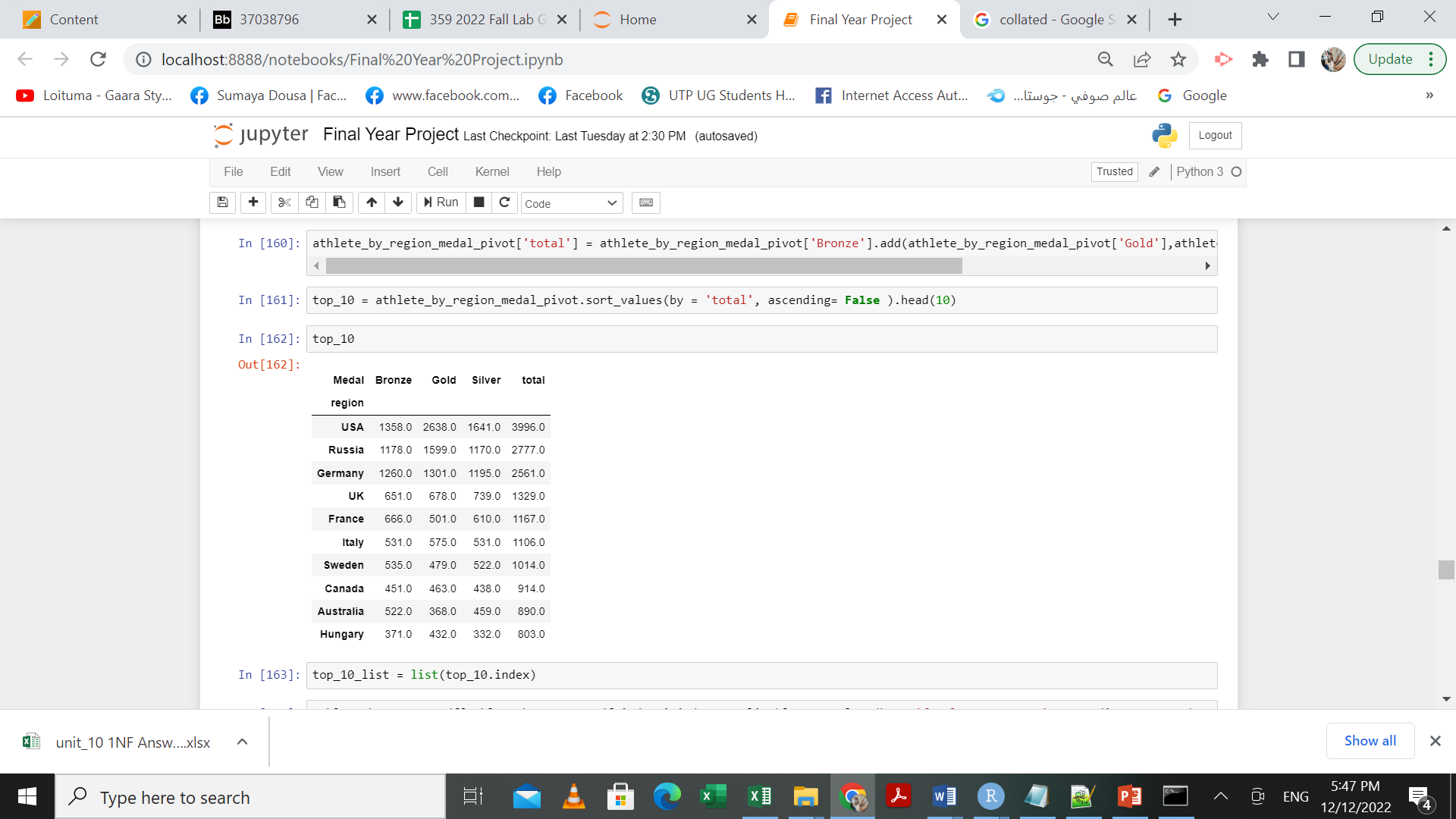
## **What are the top five countries in terms of No of medals achieved throughout the Olympics history and what is the percentage of female representation in these teams?**

Unit(s): Medal/Country

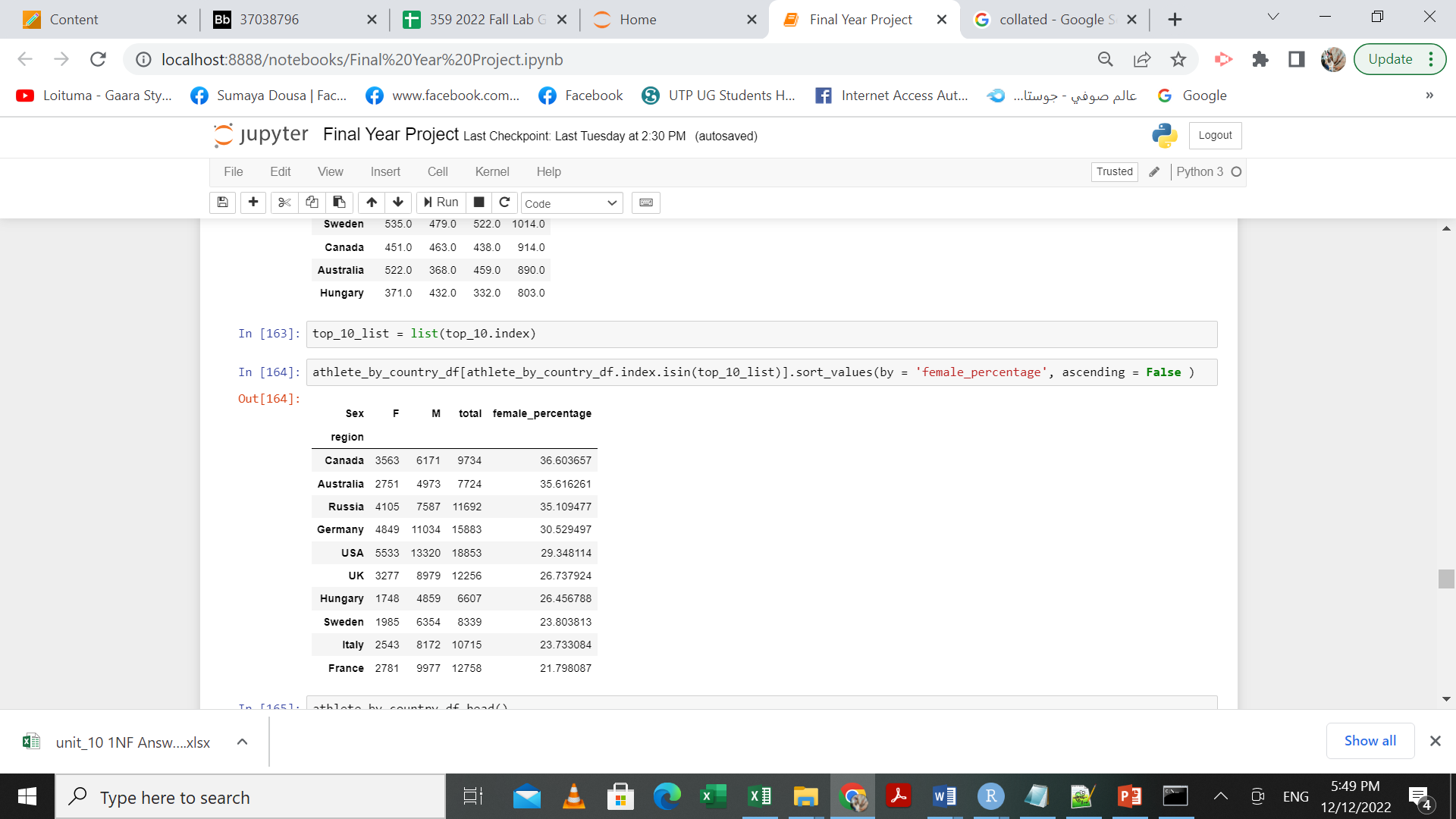
To answer this question, the Region and the Medal fields were used.. The medal column indicates the type of medal the athlete achieved on that year. Nan indicates that they have not received any.



Next, we aggregated the dataframe based on the column of the region and then calculated the number of medals achieved by these teams using the value\_counts() function. After that we only filter the 10 highest countries in terms of total number of medals and the results were as below:



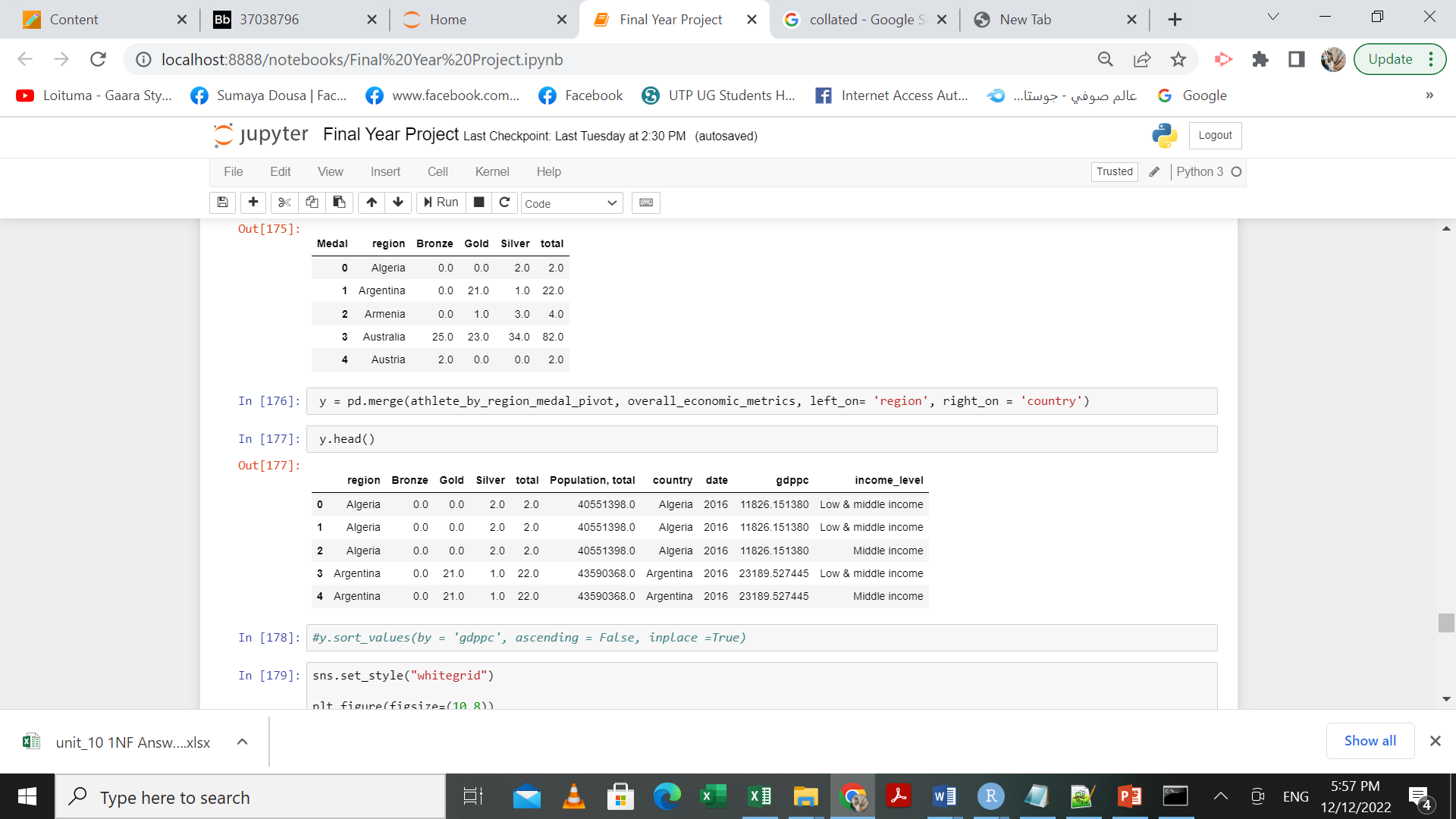
Then we looked at the female representation on these country using the output of Q1 and the result were as follows:



## **Does country GDP and expenditure on sport influence the number of medals achieved? Is there any correlation between spending on sport and the number of medals?**

Unit(s): $/country, Medal/country

To answer this question, we merged the dataset of Olympics data history and economic metrics, after aggregate the Olympics history data by region and calculate the total number of medals achieved by that region/country



It is worth mentioning that these data represents the recent Rio 2016 Olympics only.

The fields used to answer the above question were: Region, total. GDPPC and income level.

After that we plotted those fields using seaborn package to help us answer above question.

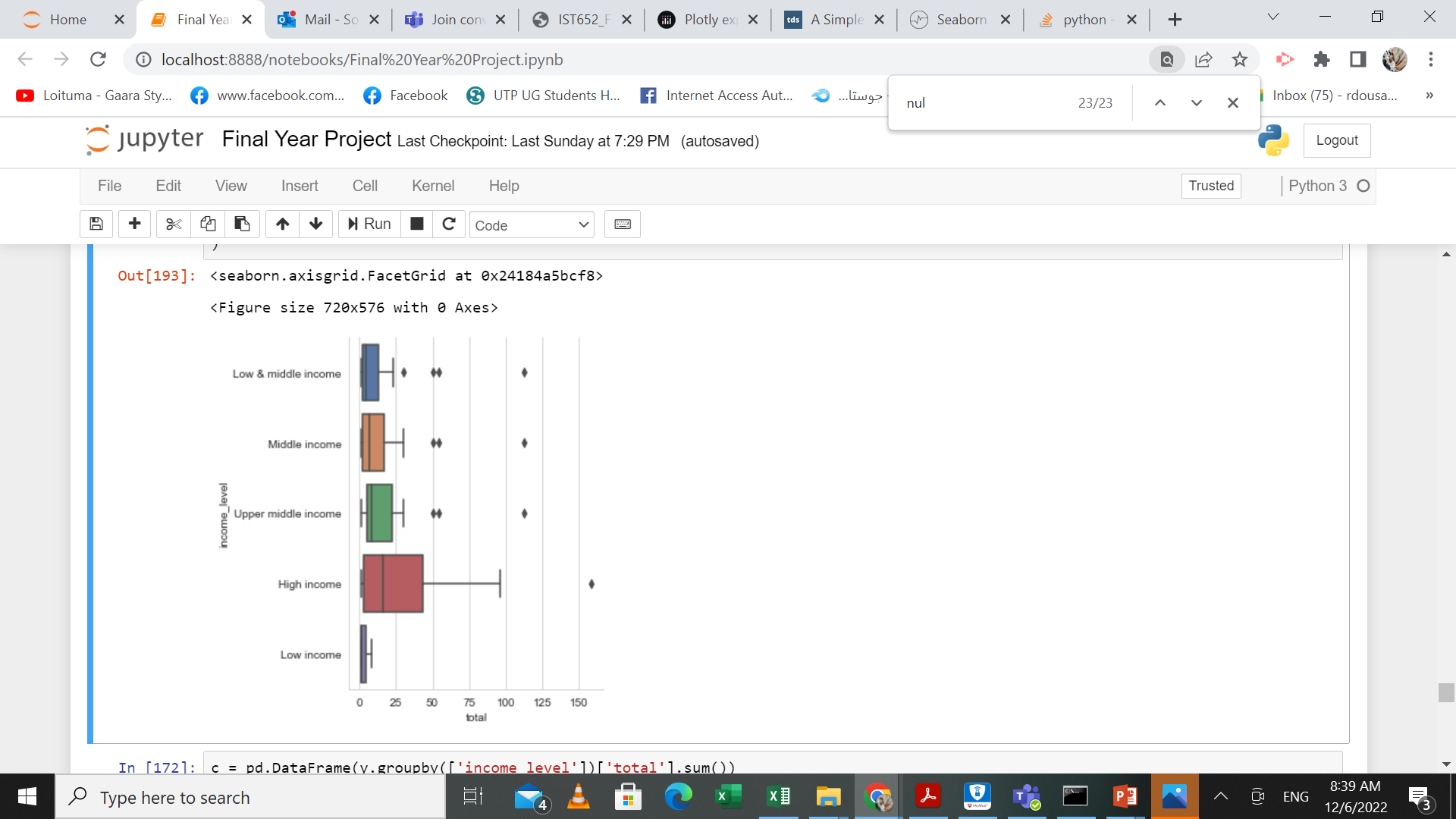


Figure Medal distribution by income level

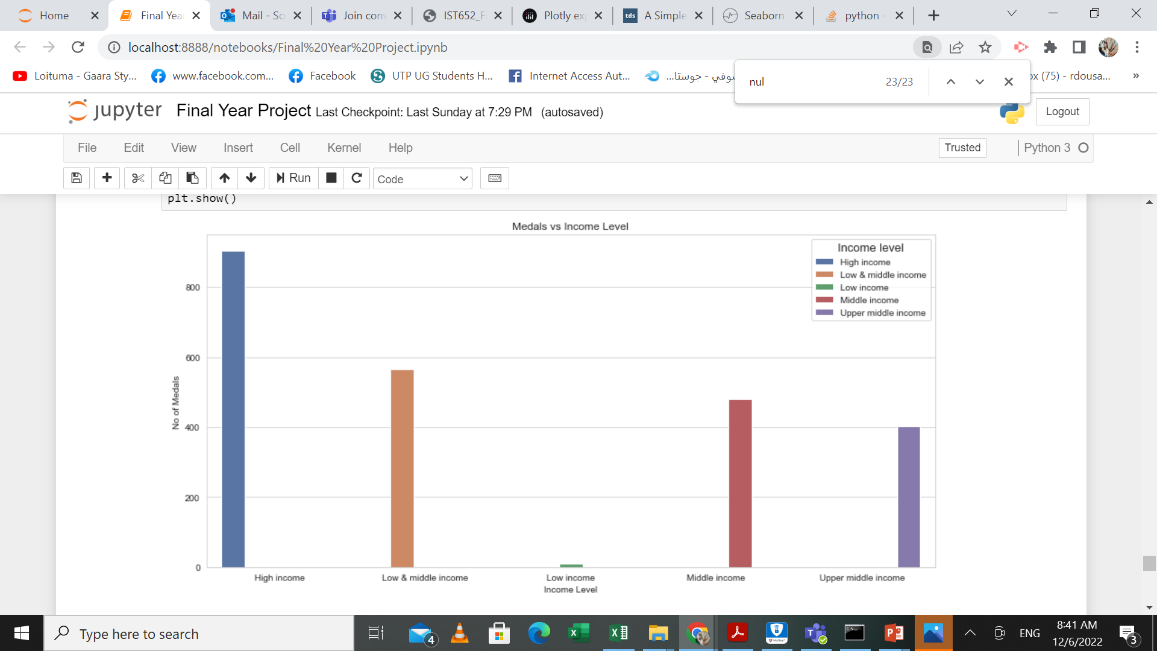


Figure 2 total number of medals by income level

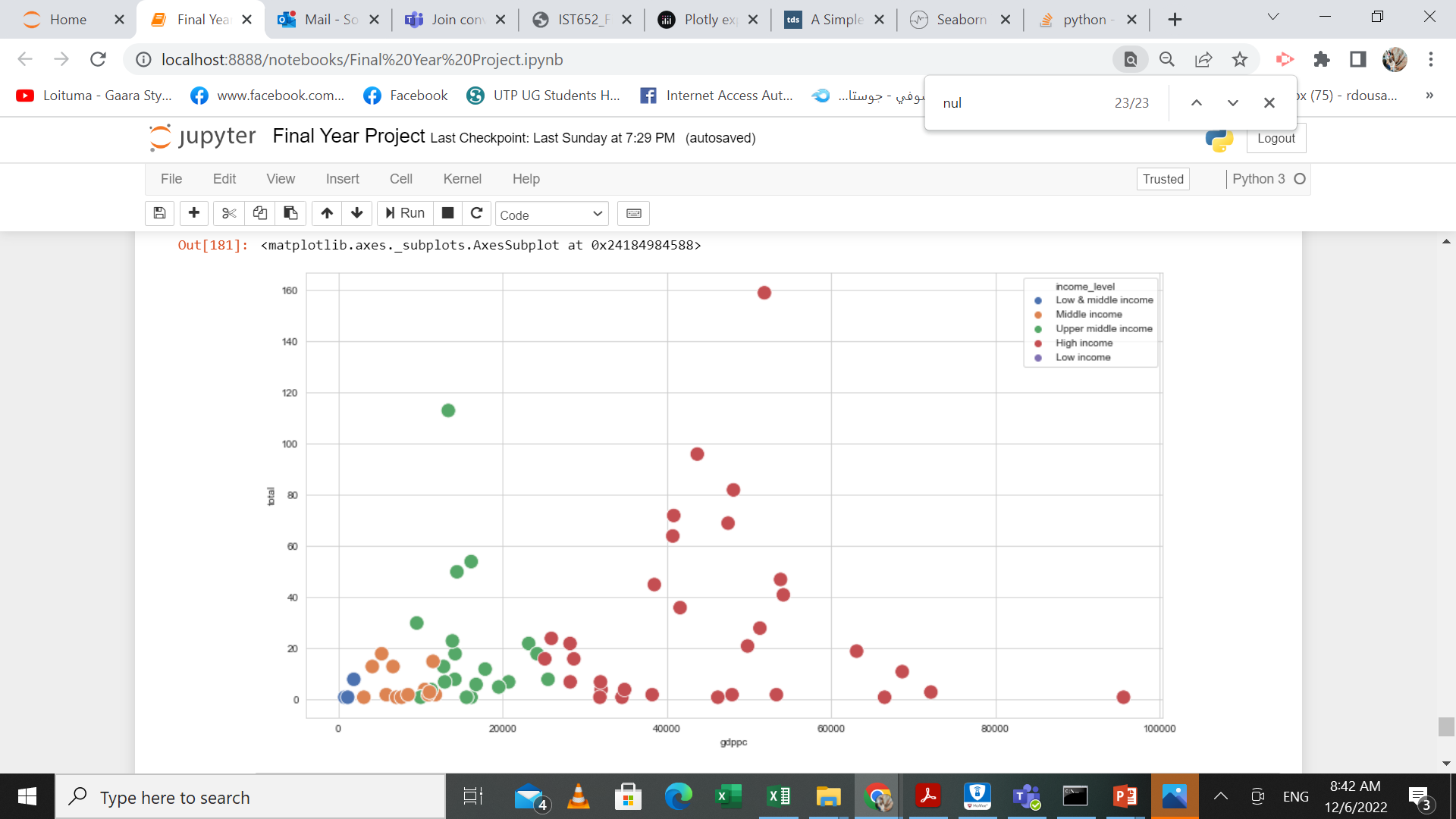


Figure 3 Correlation between GDPPC and No of medals

# **Description of Python Program:**

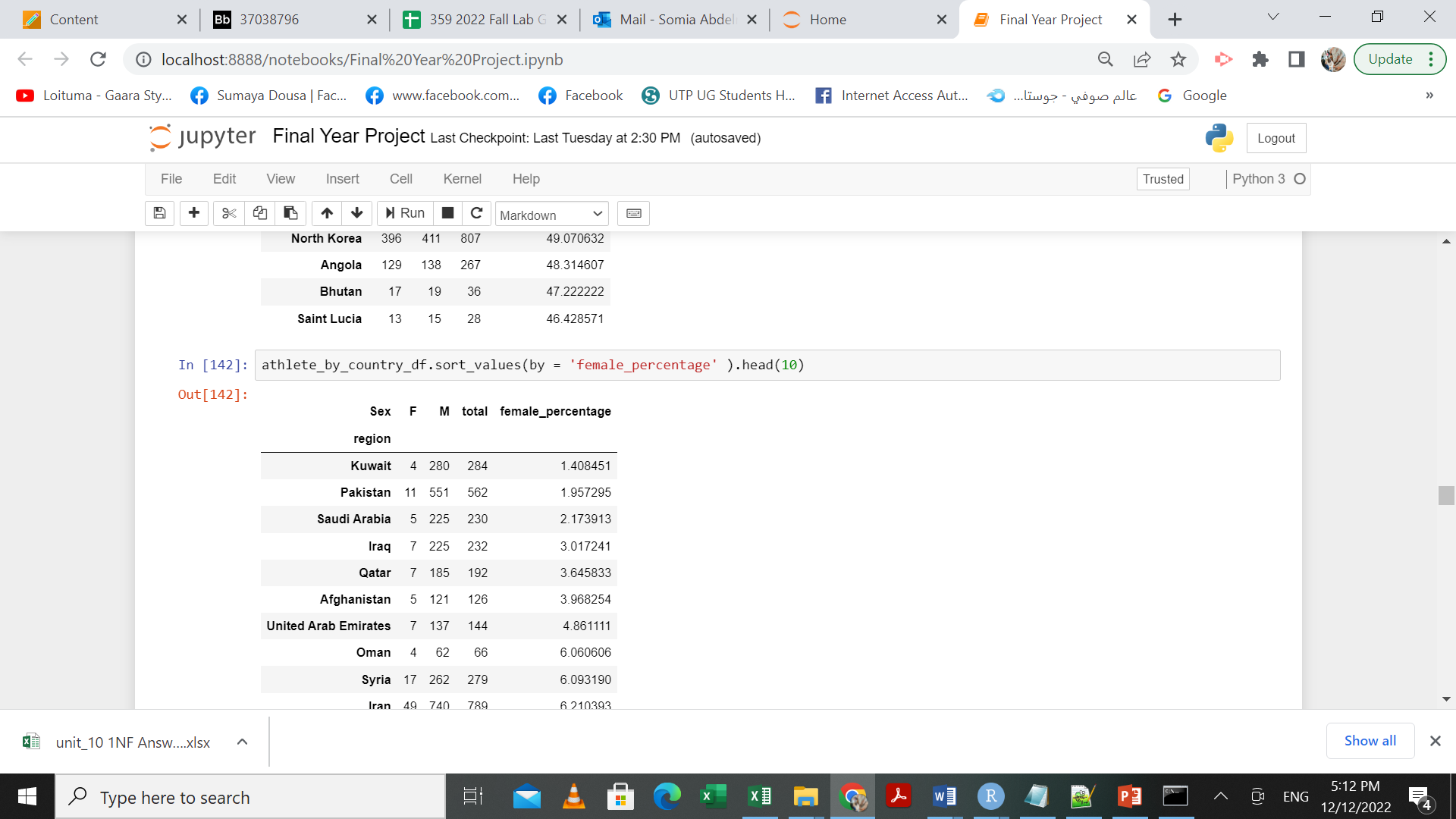
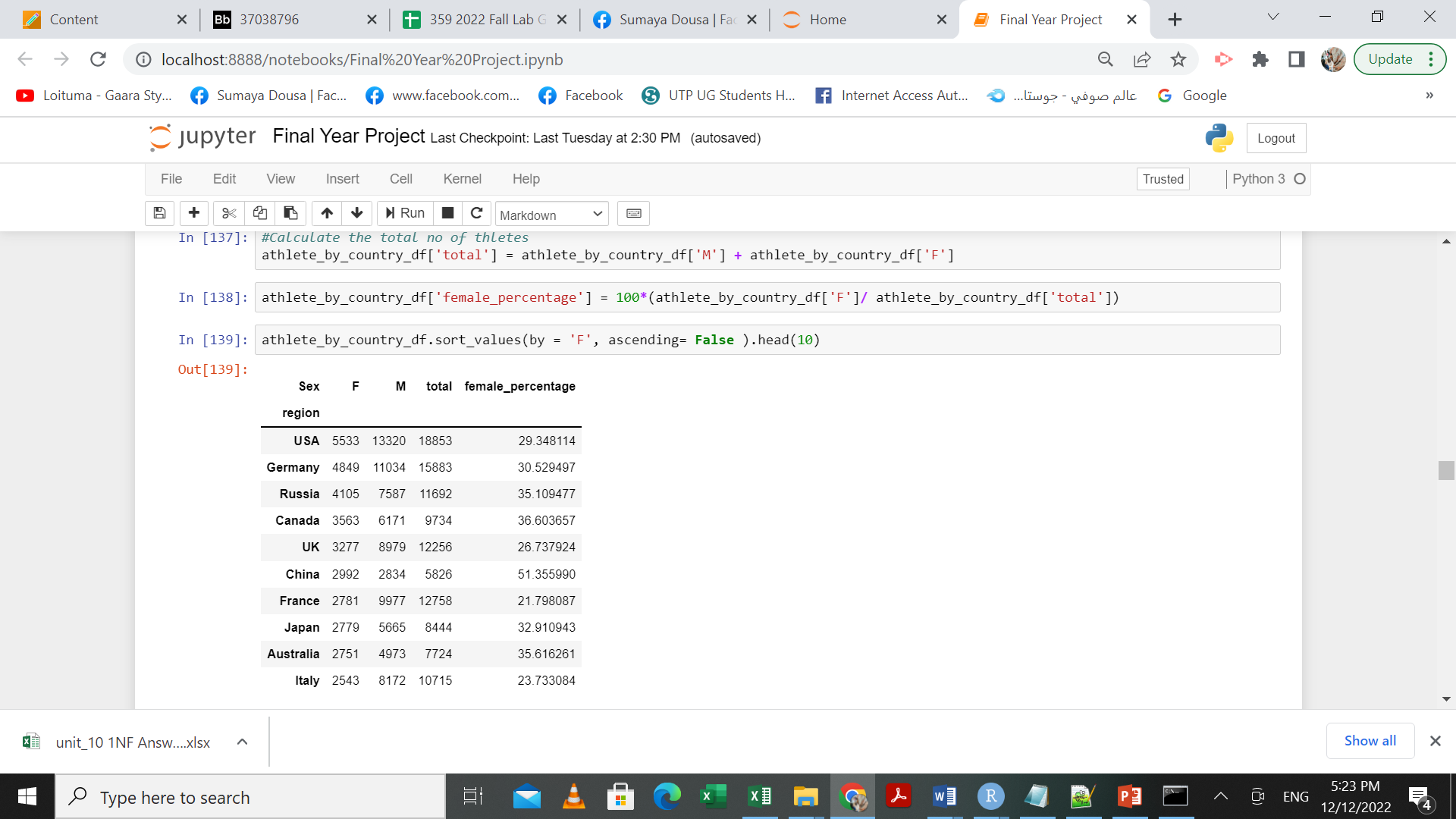
The program is written in python language in Jupyter notebook (.ipynb). Below is the summary of what the program does:

* First it imports the important packages such pandas, numpy, wbdata etc
* Then it reads the csv file into a dataframe for olympics datatset
* For economic metrics dataset, it used the wbdata package to read the metrics specified into a dataframe.
* Runs number of commands to do data exploration/data cleaning and data preprocessing
* Runs commands for analyzing the data to calculate the sum and average of certain metrics such as Total No of medals after grouping the data by certain category such as country/region.
* Stores result in new dataframes
* Visualize the outputted dataframe using seaborn visualization package.
* Exports the results into a CSV file.

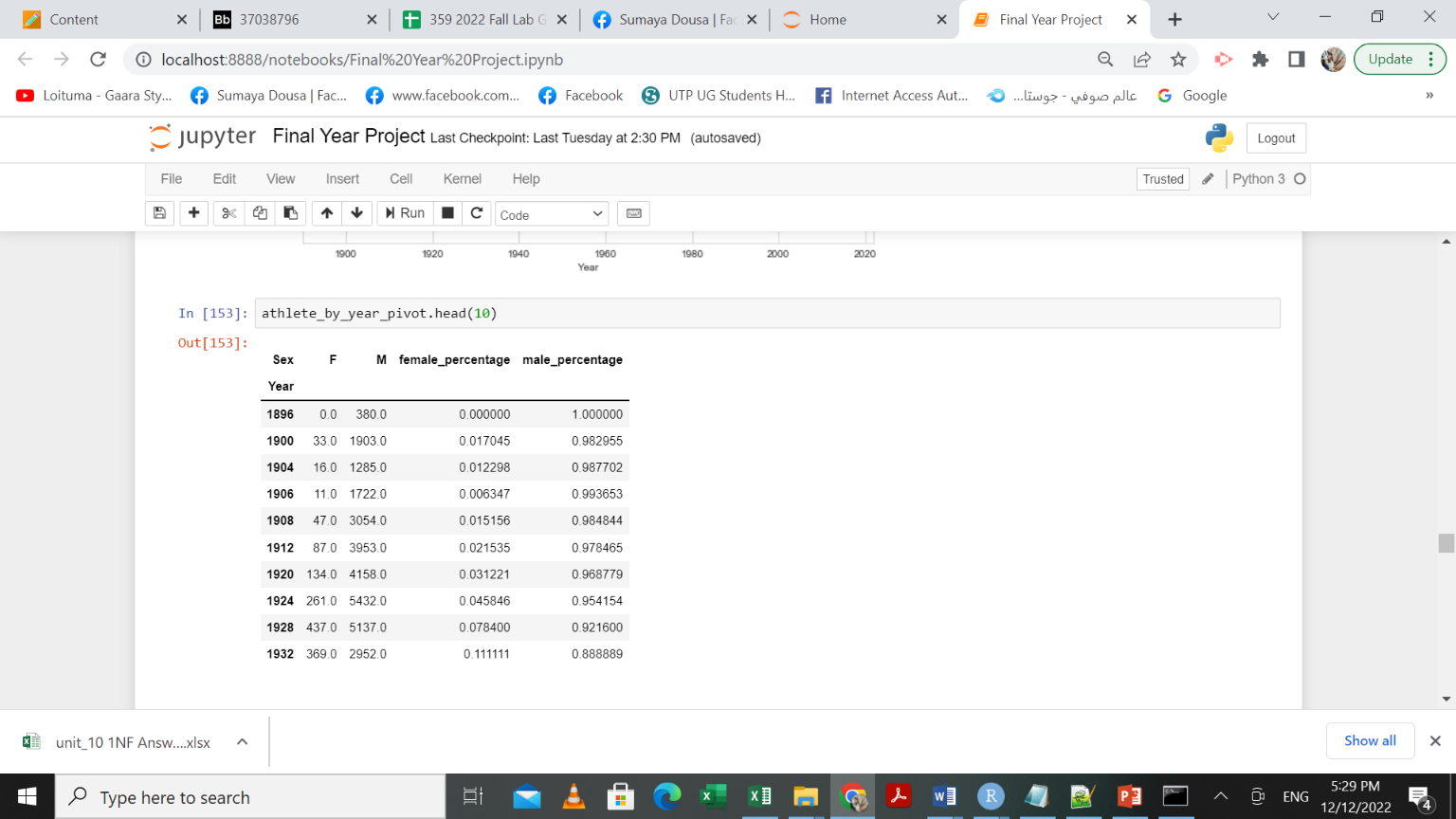
# **Documentation of the output of Python code:**

Below screenshots show the output of the python code. (dataframes and visuals)

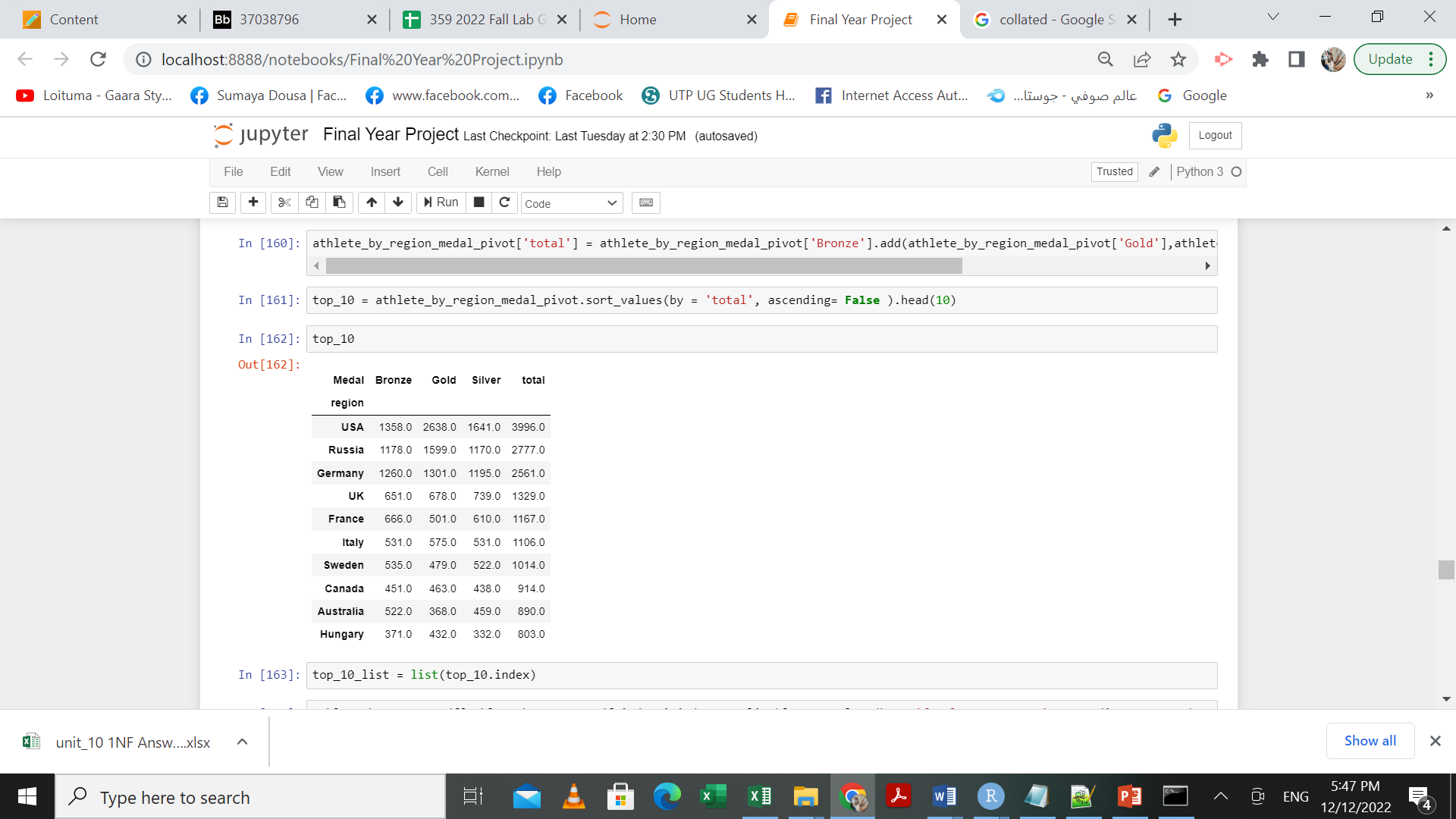
Q1:

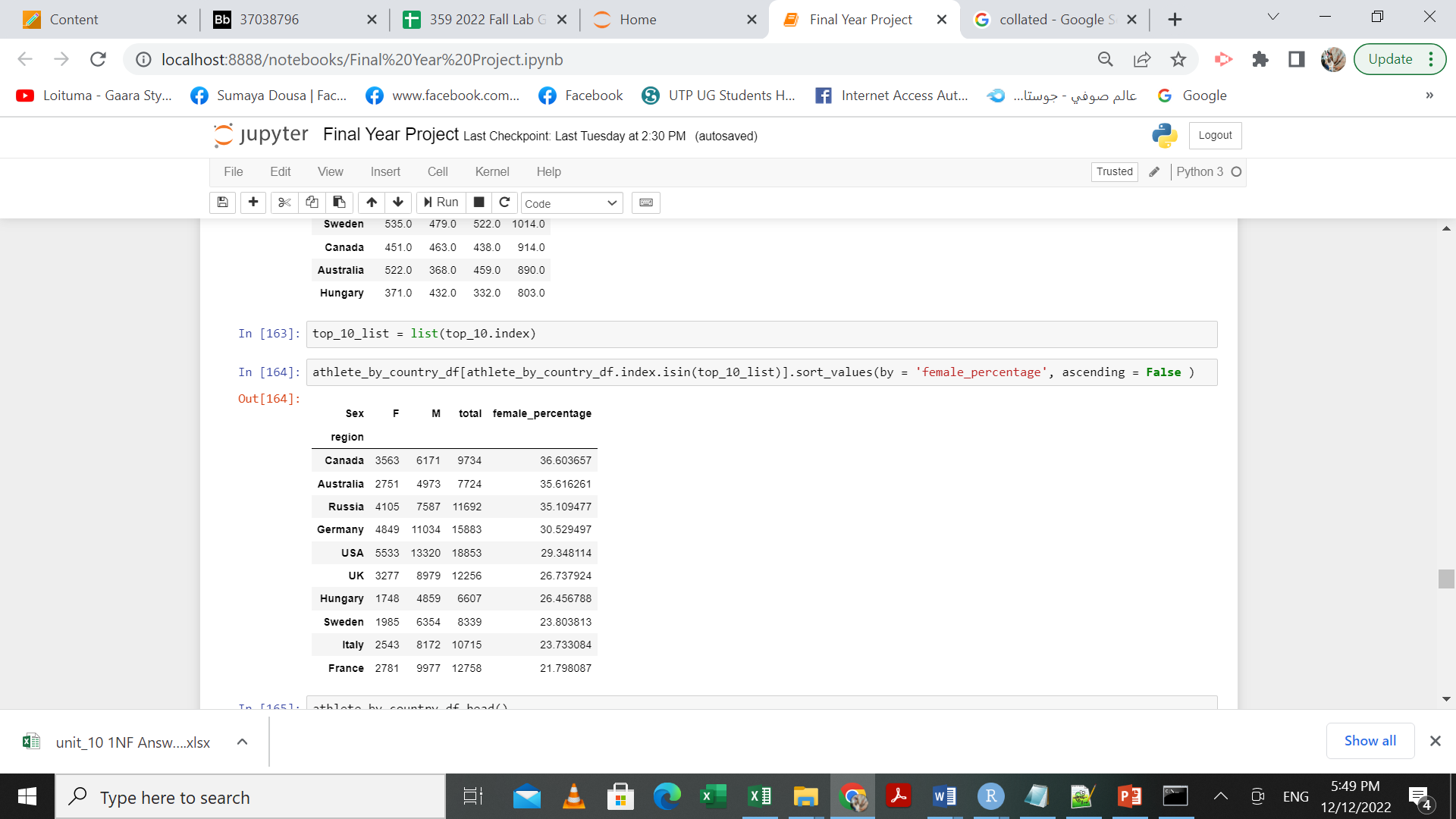


Q2:

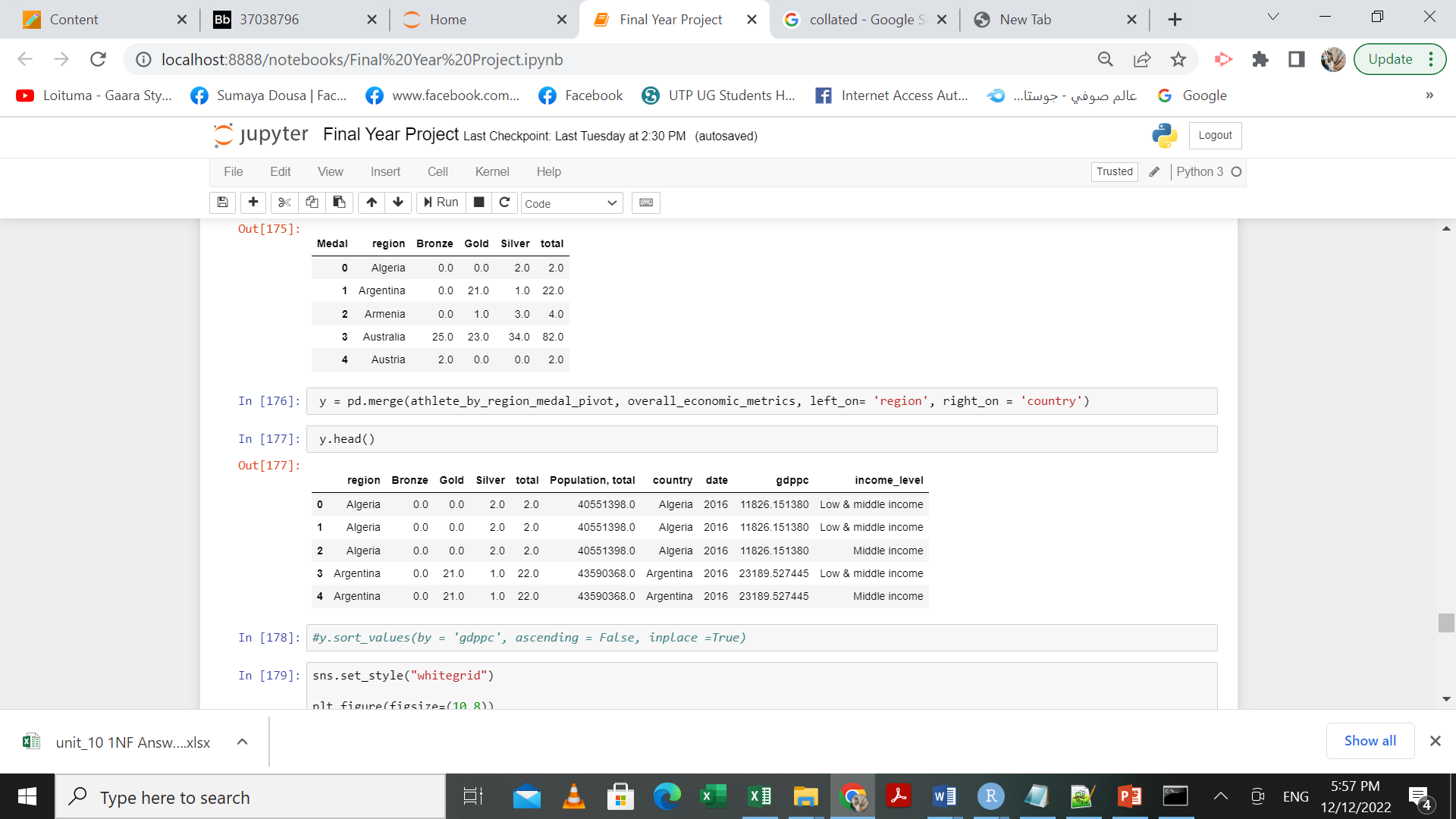


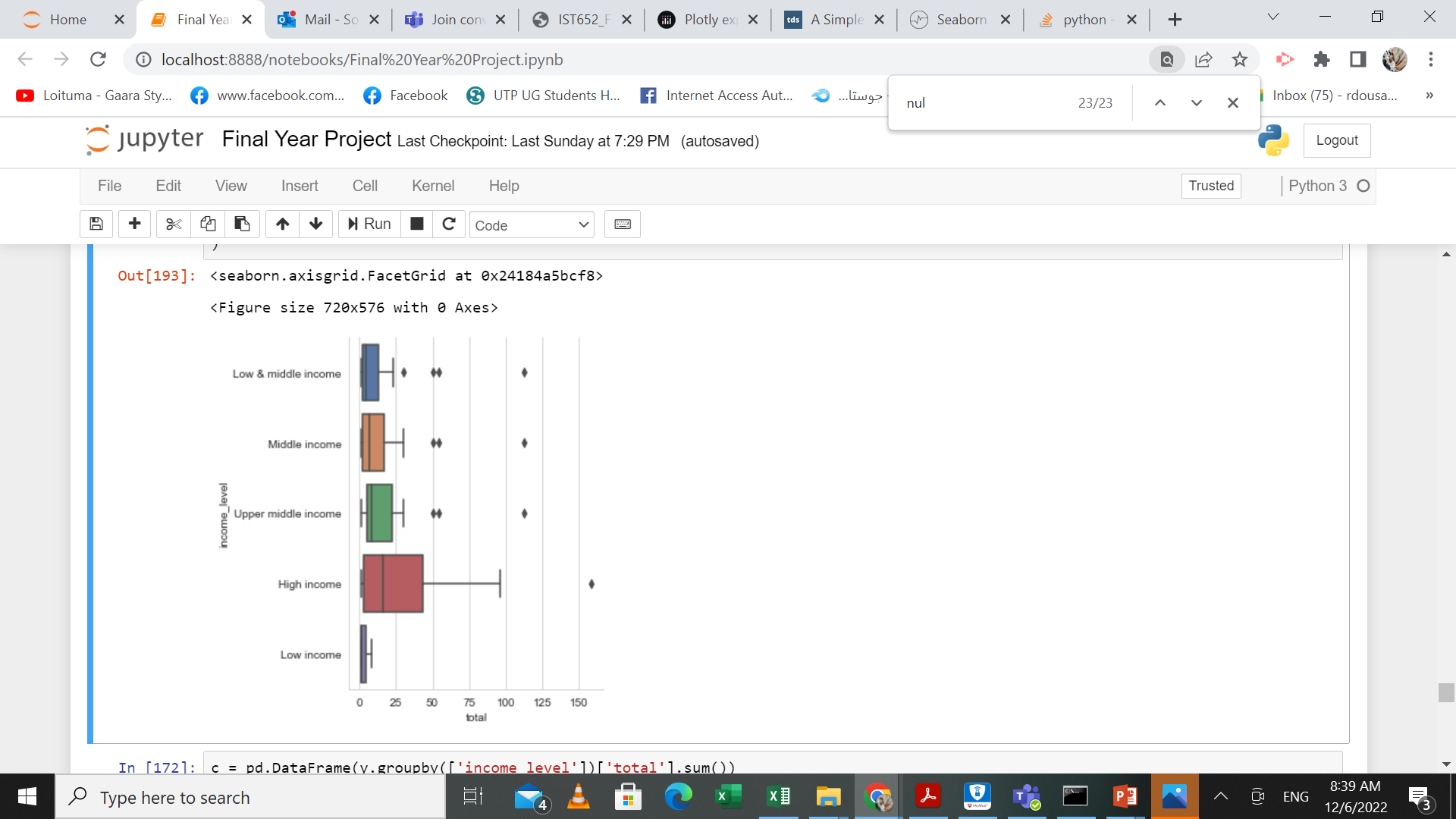
Q3:

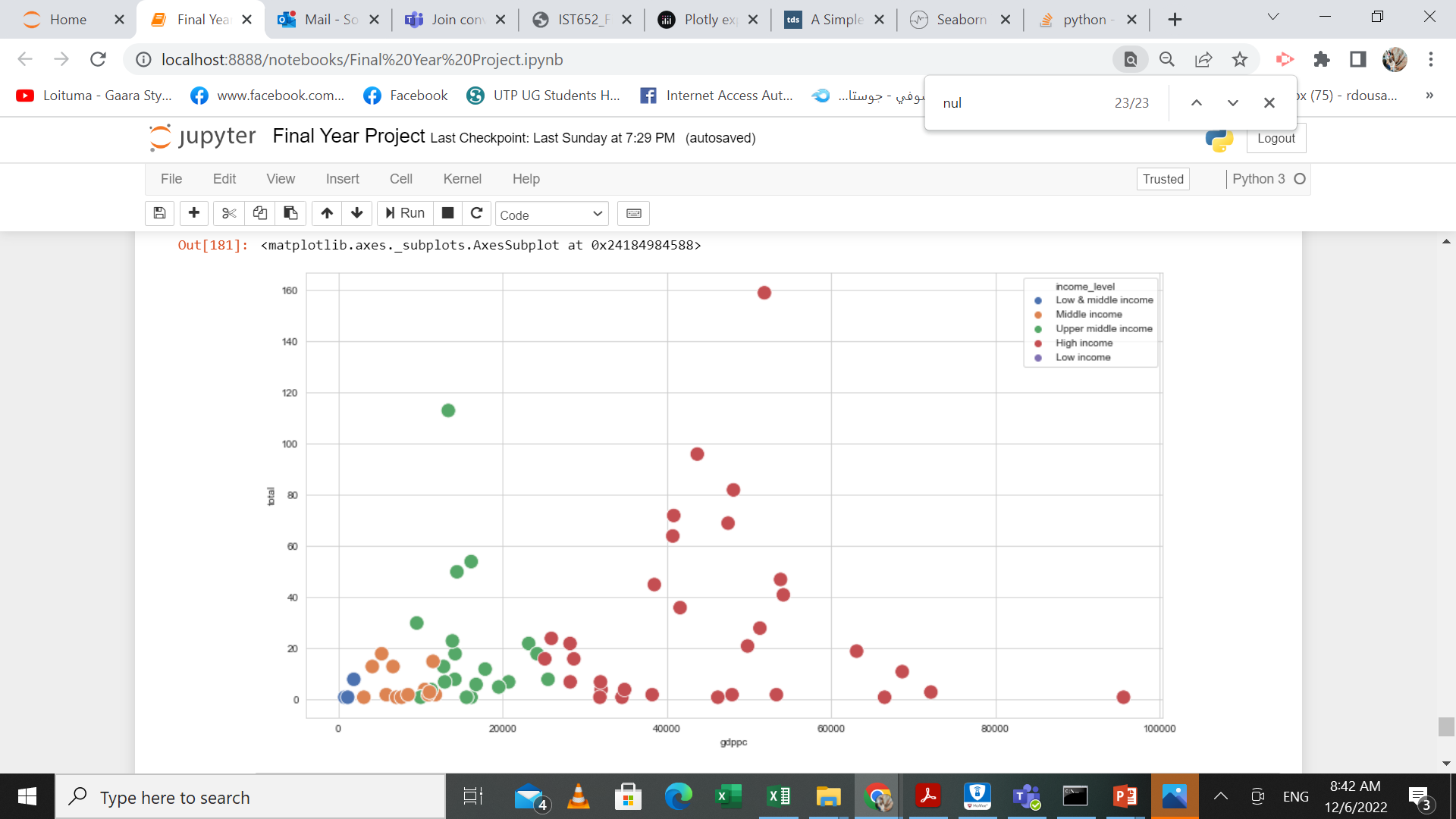




Q4:



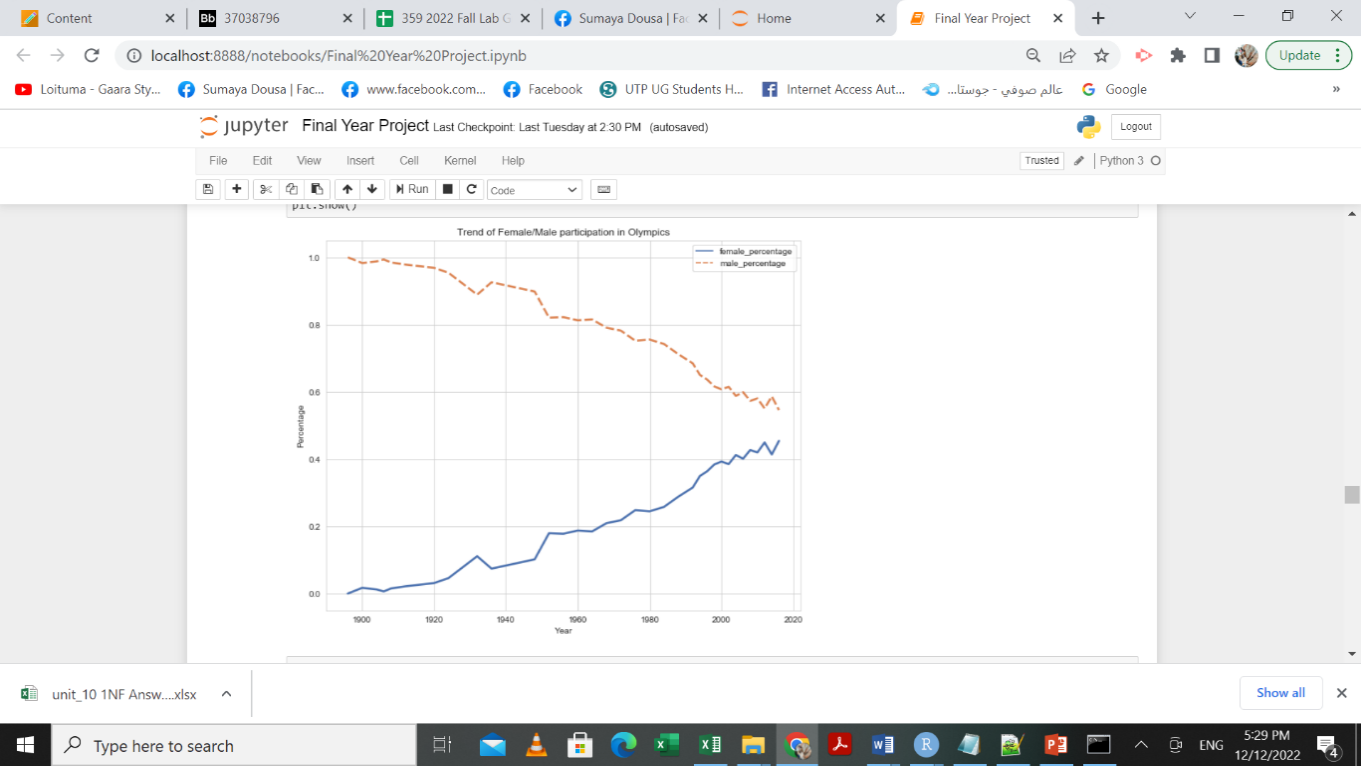
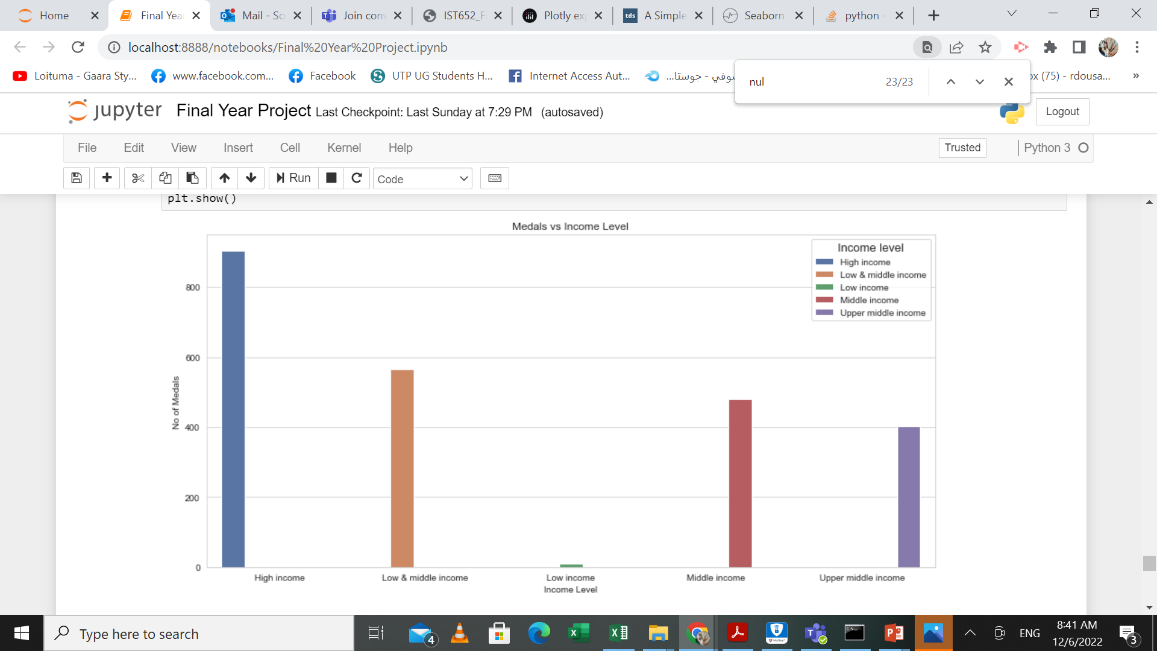




All dataframes above were stored in CSV files using the panadas to\_csv() function with different sizes and number of rows based on the unit of anlaysis used.

# **Conclusion and Interperations:**

Based on the anlaysis above, we came to the following conclusion:

* Females suffered from low presence in sport even in developed countries
* Their percence in sport however has been increasing over years and they now represent almost half of the participants in the Olympics games.
* There is a correlation between economic standing and number of medals achieved by a country.
* High income courtiers achieved more medals
* Recommendation: More funds from the World Bank to enhance the sport infrastructure in Low income countries