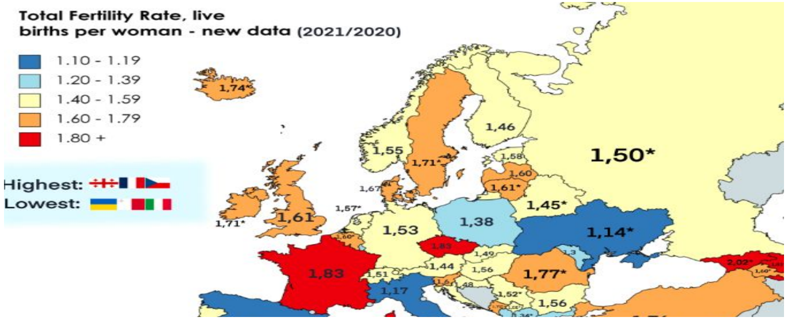
Analysis of birth rates across Europe from 2017 to 2021

Anahita Hamzeh
Fauzia Sabrina
Ilie Georgiana Nicole

Rationale

This research aims to comprehensively analyze European birth rates from 2017 to 2021, and a correlation between these birth rates.

The study relies on quantitative secondary data, which was obtained from Eurostat. With the help of Python, we can gather information and process it, displaying it in graphs to further depict and interpret the findings.



Research Question and Data collection

To what extent is the number of births determined by the GDP per capita in Europe?

The first set of data, extracted from Eurostat, was evaluated and used as the main dataset. It includes a simple data set of all European countries and the number of births during the years 2017-2021.

IMPORT DATA

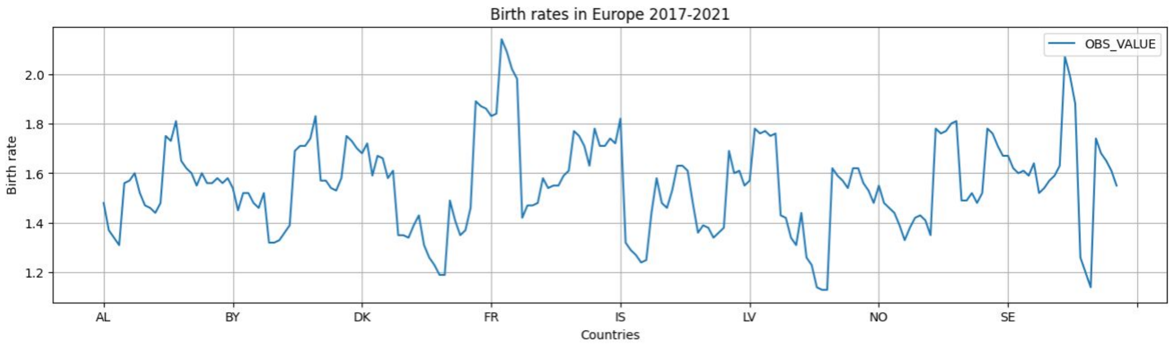
```python
import pandas as pd
from matplotlib import pyplot as plt

df = pd.read_csv("/content/drive/MyDrive/birth_rates.csv")
print(df)
```

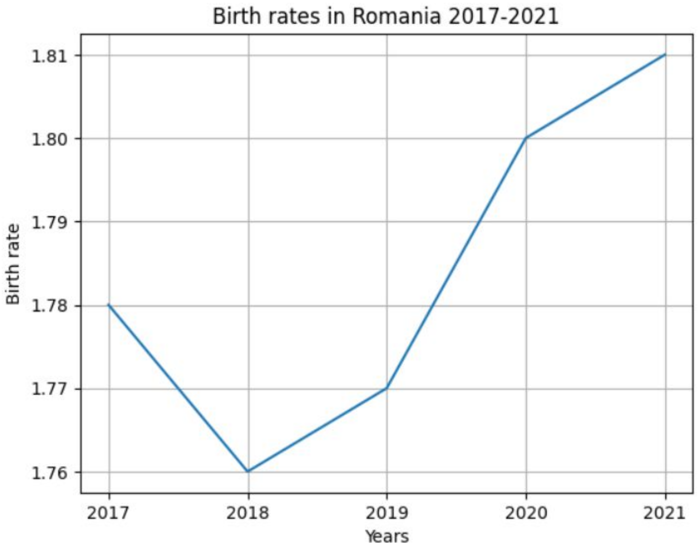
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 197 entries, 0 to 196
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   DATAFLOW     197 non-null    object
 1   LAST UPDATE  197 non-null    object
 2   freq         197 non-null    object
 3   indic_de     197 non-null    object
 4   geo          197 non-null    object
 5   TIME_PERIOD  197 non-null    int64
 6   OBS_VALUE    197 non-null    float64
 7   OBS_FLAG     19 non-null     object
dtypes: float64(1), int64(1), object(6)
memory usage: 12.4+ KB
```

Results

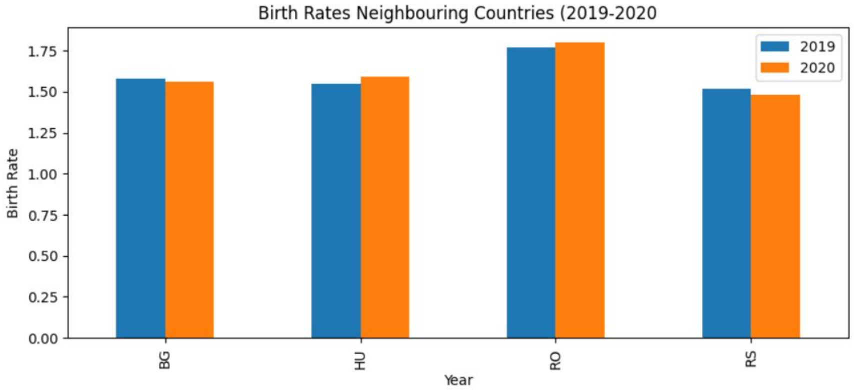
	TIME_PERIOD	OBS_VALUE
count	197.000000	197.000000
mean	2018.883249	1.552843
std	1.407550	0.190333
min	2017.000000	1.130000
25%	2018.000000	1.430000
50%	2019.000000	1.560000
75%	2020.000000	1.680000
max	2021.000000	2.140000



Results



Birth rates in Romania 2017-2021
1.81
1.80
1.79
1.78
1.77
1.76
Birth rate
2017
2018
2019
2020
2021
Years



Birth Rates Neighbouring Countries (2019-2020
2019
2020
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Birth Rate
BG
HU
RO
RS
Year

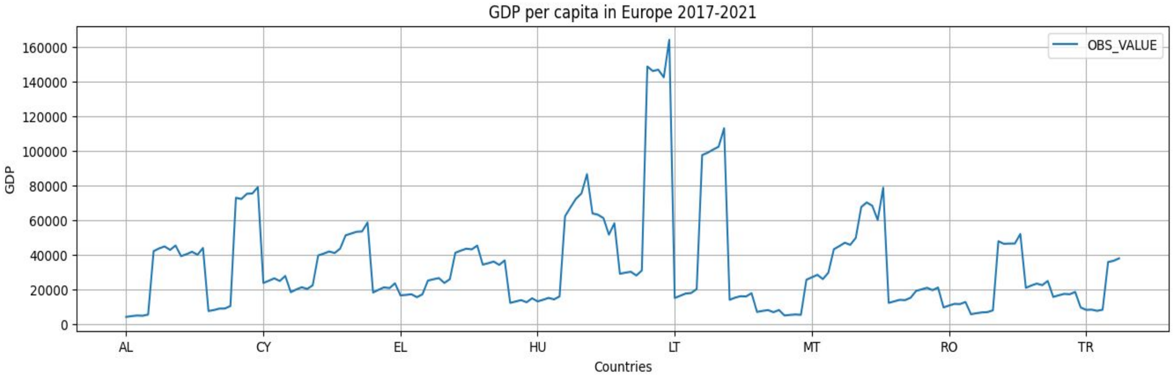
The second set of data imported

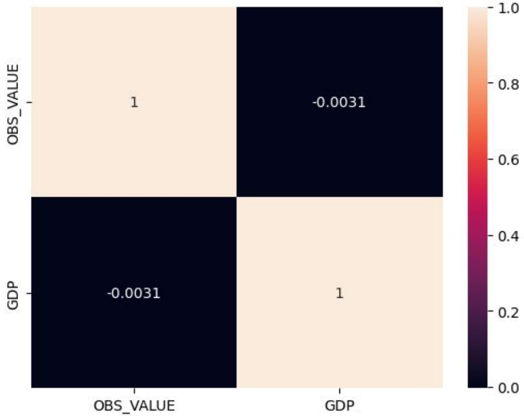
```
# importing the dataset for GDP per capita in Europe

import pandas as pd
from matplotlib import pyplot as plt

df = pd.read_csv("/content/drive/MyDrive/gdp_europe.csv")
print(df)
```

Results of the analysis






```
df_merged['OBS_VALUE'].corr(df_merg
ed['GDP'])
```

Result: **-0.003077507814769421**

Linear Regression

```
# creating x and y variables

x = df_merged['GDP'] #independent
y = df_merged['OBS_VALUE'] #dependent

linear_new = pd.DataFrame({'X': x, 'Y': y})
linear_new.head()
```

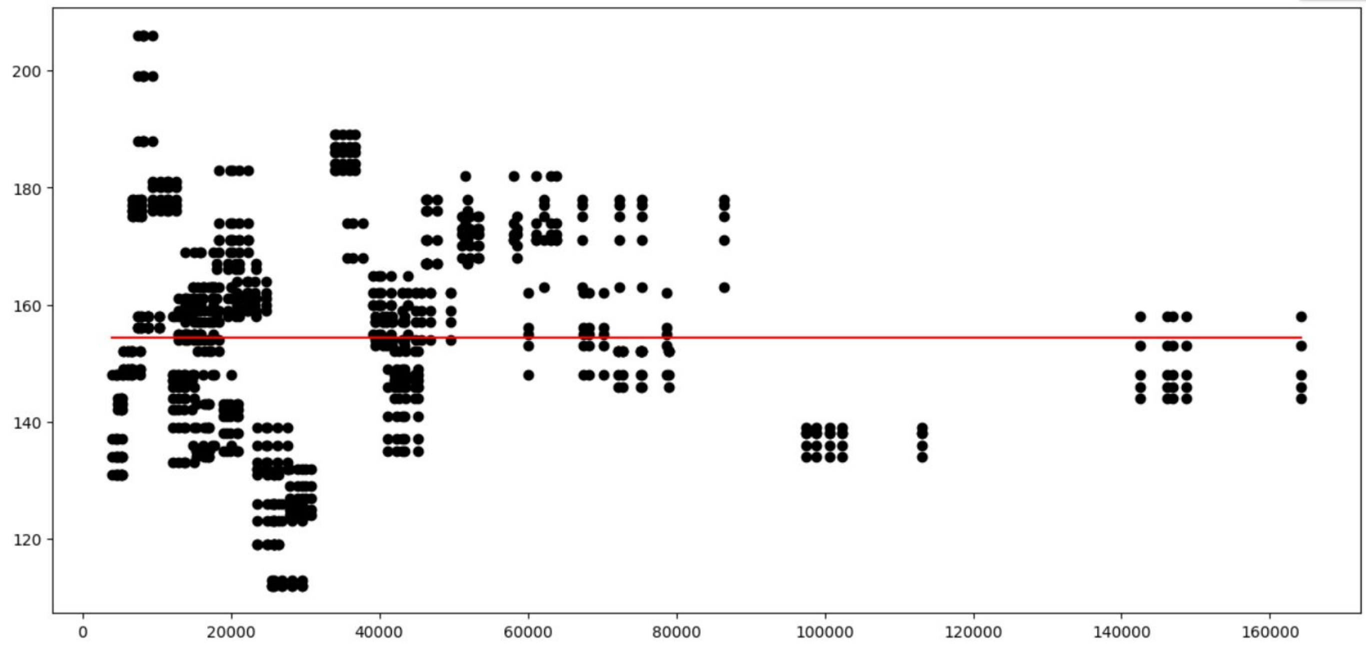
	X	Y
0	4020	148
1	4480	148
2	4820	148
3	4690	148
4	5390	148

Linear Regression

```
# plotting the data and getting a current axis of the scatter graph

import numpy as np
fig = plt.figure(figsize=(15,7))
ax = plt.gca()
ax.scatter(x, y, c ='k')
ax.plot((linear_new['X'].min(), linear_new['X'].max()),(np.mean(linear_new['Y']),
np.mean(linear_new['Y'])), color='r');
```

Results



200
180
160
140
120
0
20000
40000
60000
80000
100000
120000
140000
160000

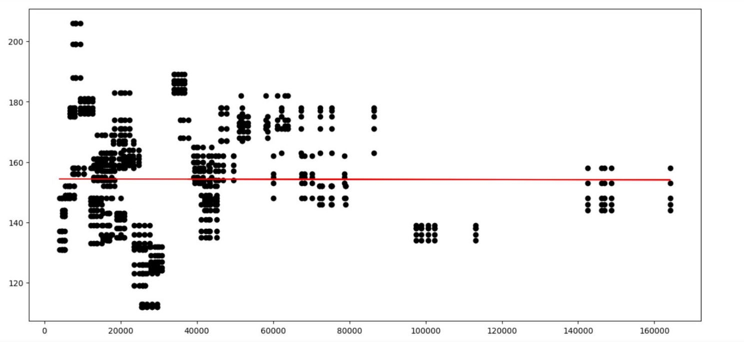
Results

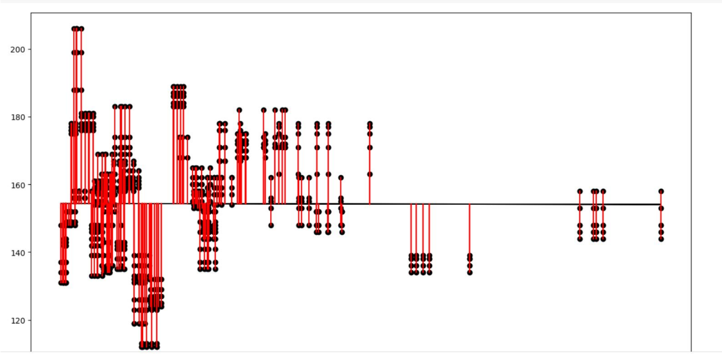
```
# finding the value that interceps with the y axis

linear_new['MeanY'] = linear_new['Y'].mean()
linear_new.head()
```

SUM OF SQUARED ERRORS = 273590.3837471783
MEAN SQUARED ERROR = 308.7927581796595
ROOT MEAN SQUARED ERROR = 17.57250005490566

	X	Y	MeanY
0	4020	148	154.325056
1	4480	148	154.325056
2	4820	148	154.325056
3	4690	148	154.325056
4	5390	148	154.325056





Interpretations (I)

Sum of Squared Errors (SSE) value, 273590.3837471783, measures the total deviation of the observed GDP per capita values from the values predicted by the model using birth rates. If GDP per capita values in the dataset would range in the millions, this SSE might suggest a relatively smaller error and more reliable model.

But if GDP per capita values are much smaller, the same SSE could indicate a lot of error and less confidence in your model.

In the case of Mean Squared Errors of 308.7927581796595 for predicting GDP per capita based on birth rates means that, on average, the model's predictions deviate from the actual data by the square root of this value, given that the error distribution is normal.

An MSE of 308.7927581796595 implies the Root Mean Squared Error (RMSE) - which indicates the standard deviation of the residuals and is often more interpretable as it is in the same units as the outcome - is about 17.57 (square root of the MSE). This means that your model's predictions are, on average, about 17.57 units (of GDP per capita) away from the observed data.

Interpretations (II)

Root Mean Squared Error (RMSE) value is 17.57250005490566. The RMSE is a measure of the average deviation of the predictions from the observed values in your dataset, also known as the prediction error.

Conclusion

Based on the results provided including SSE, MSE, and RMSE but not including any coefficients or other specifics about the model, we know that the linear regression model predicting GDP per capita based on birth rates has a certain degree of error. This error is particularly encapsulated in the RMSE value of 17.57 units. This tells us that, on average, the model's predictions of GDP per capita are approximately 17.57 units away from the actual observed values.

Based solely on the results provided so far, aside from providing a broad measure of the model's accuracy, we can't yet make any specific claims about the relationship between GDP per capita and birth rates. A more thorough analysis of your regression output would be necessary.

In addition, interpreting such results should always be done in conjunction with a residuals analysis, cross-validation, or out-of-sample testing, all of which provide important context and validity checks for your model.