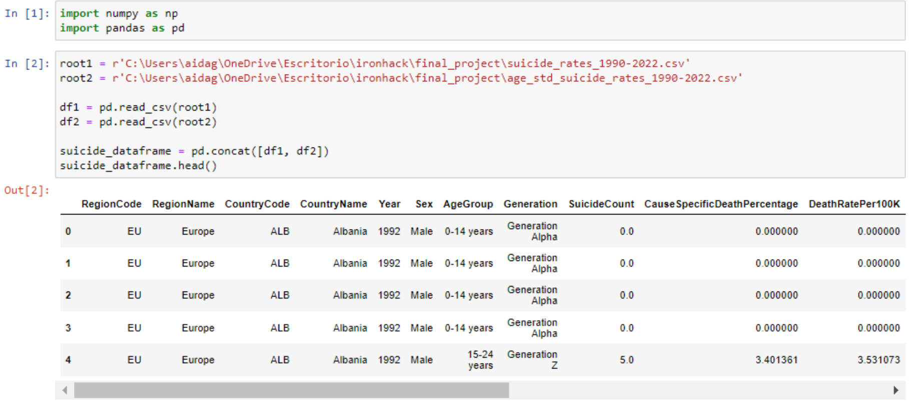
INTRODUCTION

My final course project focuses on suicide in Spain. I extracted data from a dataset covering various information from 1990 to 2021, which includes data from all countries worldwide. The reason I chose this topic is because I thought it would be a good idea to analyze a current issue that might be of interest to a ministry of the Spanish government or to psychologists involved in treating suicide cases.



```
In [1]: import numpy as np
        import pandas as pd
```

```
In [2]: root1 = r'C:\Users\aidag\OneDrive\Escritorio\ironhack\final_project\suicide_rates_1990-2022.csv'
        root2 = r'C:\Users\aidag\OneDrive\Escritorio\ironhack\final_project\age_std_suicide_rates_1990-2022.csv'

        df1 = pd.read_csv(root1)
        df2 = pd.read_csv(root2)

        suicide_dataframe = pd.concat([df1, df2])
        suicide_dataframe.head()
```

Out[2]:

	RegionCode	RegionName	CountryCode	CountryName	Year	Sex	AgeGroup	Generation	SuicideCount	CauseSpecificDeathPercentage	DeathRatePer100K
0	EU	Europe	ALB	Albania	1992	Male	0-14 years	Generation Alpha	0.0	0.000000	0.000000
1	EU	Europe	ALB	Albania	1992	Male	0-14 years	Generation Alpha	0.0	0.000000	0.000000
2	EU	Europe	ALB	Albania	1992	Male	0-14 years	Generation Alpha	0.0	0.000000	0.000000
3	EU	Europe	ALB	Albania	1992	Male	0-14 years	Generation Alpha	0.0	0.000000	0.000000
4	EU	Europe	ALB	Albania	1992	Male	15-24 years	Generation Z	5.0	3.401361	3.531073

OBSERVATIONS AND DATA CLEANING

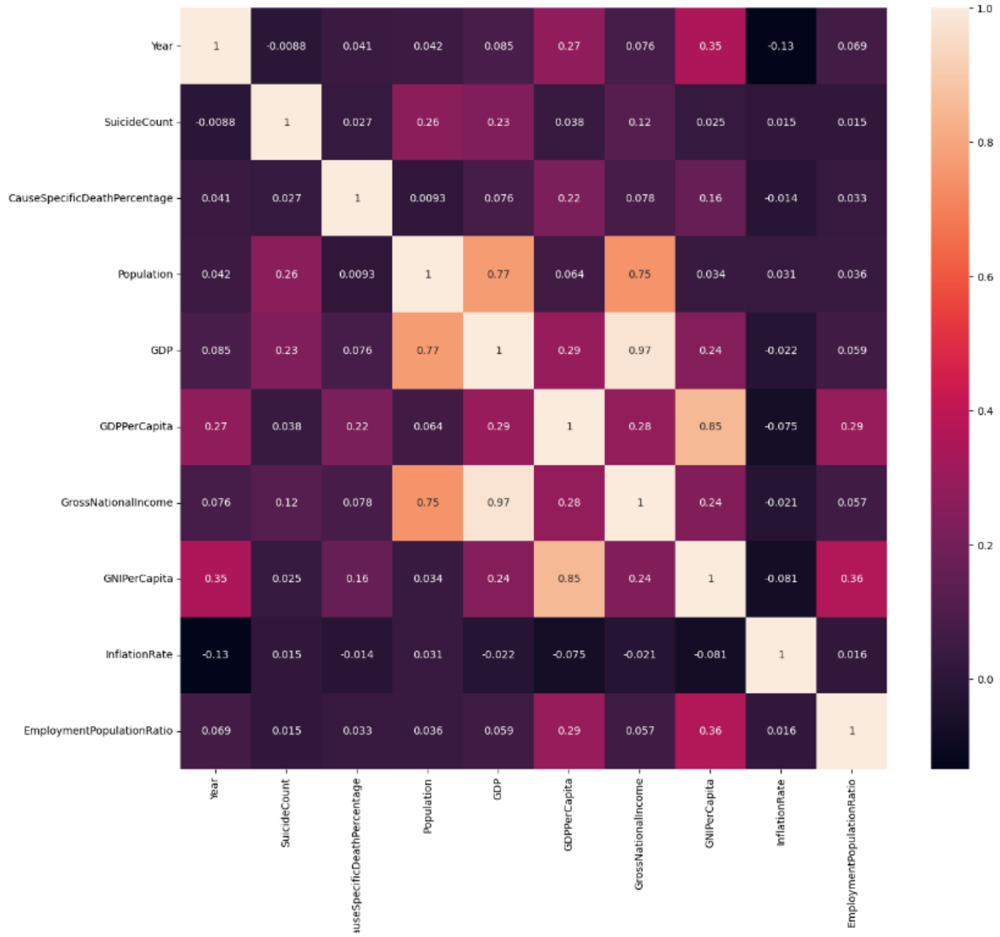
First, starting from a dataset, I have worked on observing the information it contains and performing cleaning tasks by removing columns that I consider did not contribute much to the dataset or that those columns could be redundant, such as 'GNI', since there is another column called 'GNI per capita', 'DeathRatePer100k', and 'STD death rate'.

For future operations, I have stored the columns in a numerical and a categorical variable.

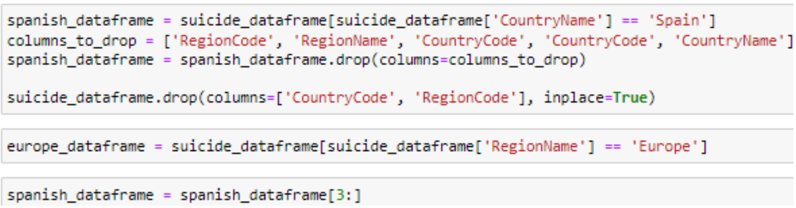
Regarding the null values that I found, I replaced the NaN values in the numerical columns with a calculation of the mean of those columns, as it was giving an error and emptying the dataset every time I tried to use .dropna. For the categorical columns with NaNs, I decided to use "No info".

EDA

Looking at the graph of the correlation matrix, I have eliminated the 'GDP' column because it has a 0.97 hillality, it drew attention to lower values and that is a very high number, apart from that column not contributing much to the problem of suicide.



Then, I filtered the data to something smaller to make it more manageable and focus on suicide in Spain. However, I also filtered the data to obtain information about the issue in Europe and make some comparisons with Spain.



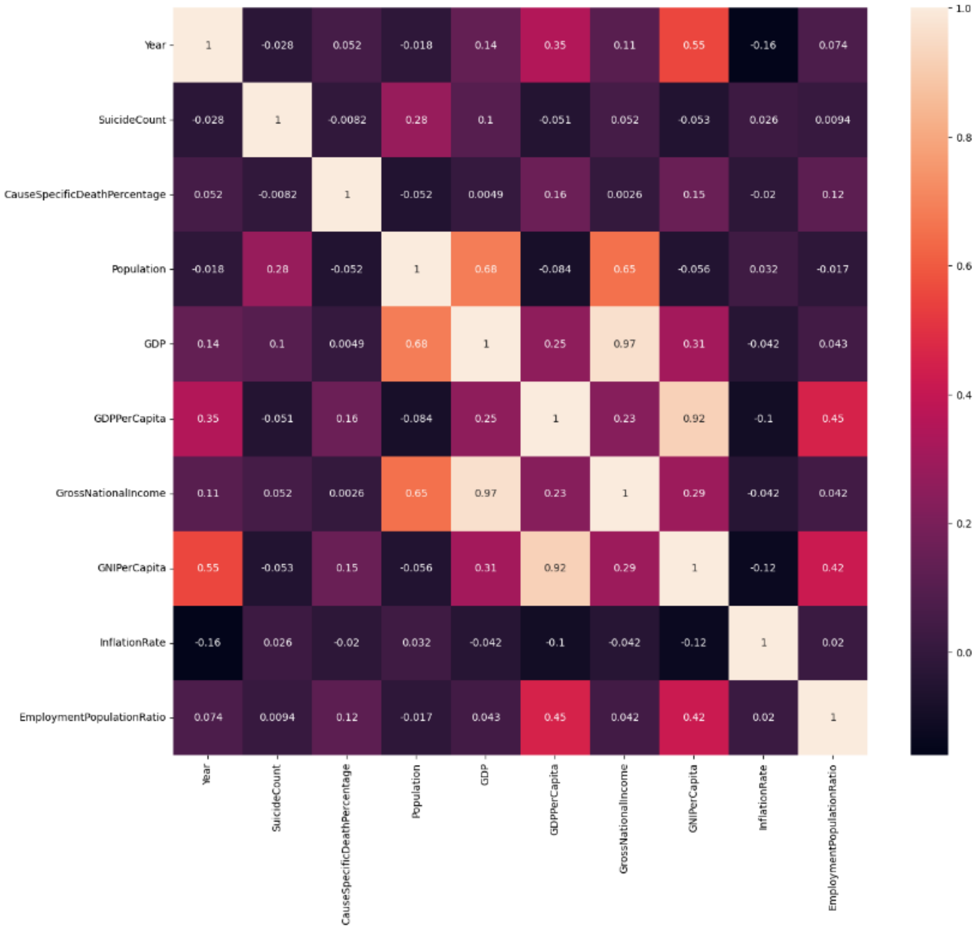
```
spanish_dataframe = suicide_dataframe[suicide_dataframe['CountryName'] == 'Spain']
columns_to_drop = ['RegionCode', 'RegionName', 'CountryCode', 'CountryCode', 'CountryName']
spanish_dataframe = spanish_dataframe.drop(columns=columns_to_drop)

suicide_dataframe.drop(columns=['CountryCode', 'RegionCode'], inplace=True)
```

```
europe_dataframe = suicide_dataframe[suicide_dataframe['RegionName'] == 'Europe']
```

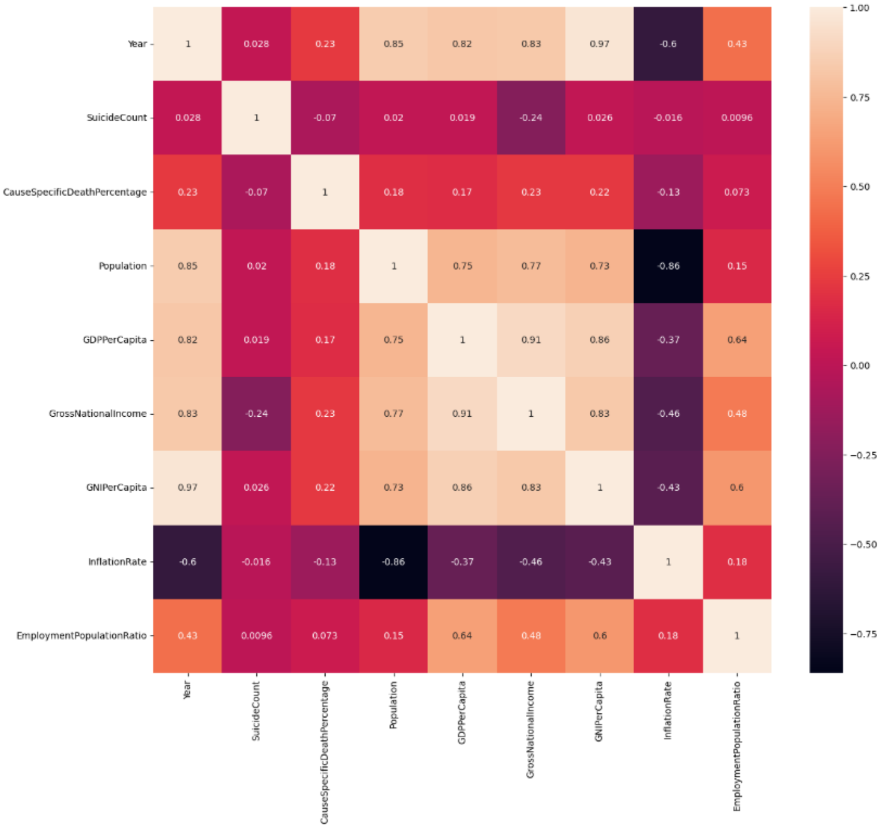
```
spanish_dataframe = spanish_dataframe[3:]
```

I noticed that the first four rows are repeated, so I filtered the Spain dataset with [3:] to show everything from that row onwards. Next, I plotted the correlation matrix graph for Europe, which apparently looks fine.

However, after plotting the correlation matrix specifically for Spain, I observed high collinearity, but all columns are important for analyzing the issue, so I haven't removed any

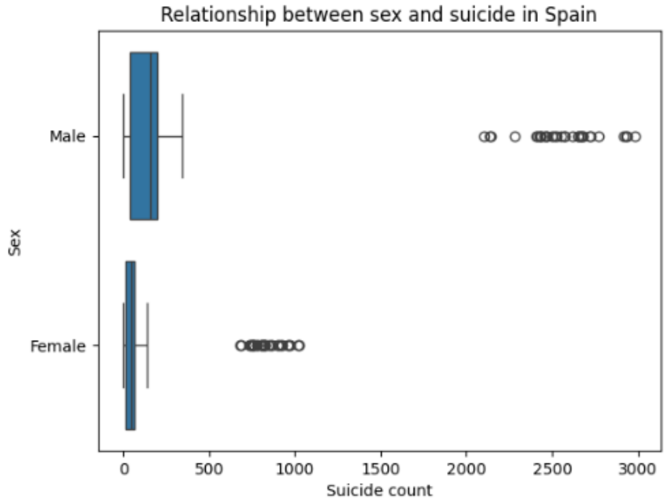

Year 1 -0.028 0.052 -0.018 0.14 0.35 0.11 0.55 -0.16 0.074
SuicideCount -0.028 1 -0.0082 0.28 0.1 -0.051 0.052 -0.053 0.026 0.0094
CauseSpecificDeathPercentage 0.052 -0.0082 1 -0.052 0.0049 0.16 0.0026 0.15 -0.02 0.12
Population -0.018 0.28 -0.052 1 0.68 -0.084 0.65 -0.056 0.032 -0.017
GDP 0.14 0.1 0.0049 0.68 1 0.25 0.97 0.31 -0.042 0.043
GDPPerCapita 0.35 -0.051 0.16 -0.084 0.25 1 0.23 0.92 -0.1 0.45
GrossNationalIncome 0.11 0.052 0.0026 0.65 0.97 0.23 1 0.29 -0.042 0.042
GNIPerCapita 0.55 -0.053 0.15 -0.056 0.31 0.92 0.29 1 -0.12 0.42
InflationRate -0.16 0.026 -0.02 0.032 -0.042 -0.1 -0.042 -0.12 1 0.02
EmploymentPopulationRatio 0.074 0.0094 0.12 -0.017 0.043 0.45 0.042 0.42 0.02 1
Year
SuicideCount
CauseSpecificDeathPercentage
Population
GDP
GDPPerCapita
GrossNationalIncome
GNIPerCapita
InflationRate
EmploymentPopulationRatio
1.0
0.8
0.6
0.4
0.2
0.0

column.

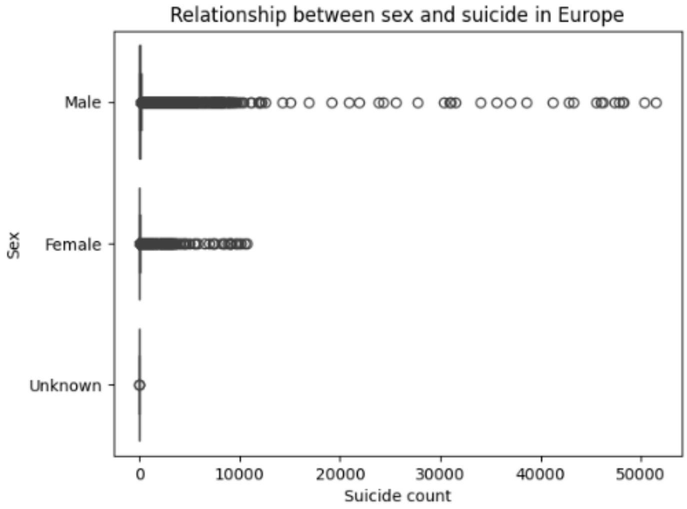

Year 1 0.028 0.23 0.85 0.82 0.83 0.97 -0.6 0.43
SuicideCount 0.028 1 -0.07 0.02 0.019 -0.24 0.026 -0.016 0.0096
CauseSpecificDeathPercentage 0.23 -0.07 1 0.18 0.17 0.23 0.22 -0.13 0.073
Population 0.85 0.02 0.18 1 0.75 0.77 0.73 -0.86 0.15
GDPPerCapita 0.82 0.019 0.17 0.75 1 0.91 0.86 -0.37 0.64
GrossNationalIncome 0.83 -0.24 0.23 0.77 0.91 1 0.83 -0.46 0.48
GNIPerCapita 0.97 0.026 0.22 0.73 0.86 0.83 1 -0.43 0.6
InflationRate -0.6 -0.016 -0.13 -0.86 -0.37 -0.46 -0.43 1 0.18
EmploymentPopulationRatio 0.43 0.0096 0.073 0.15 0.64 0.48 0.6 0.18 1
Year
SuicideCount
CauseSpecificDeathPercentage
Population
GDPPerCapita
GrossNationalIncome
GNIPerCapita
InflationRate
EmploymentPopulationRatio
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

I've also used "describe" with Transpose to see the mean, mode, quartiles, etc. Finally, I've created different types of graphs to observe before building the model.

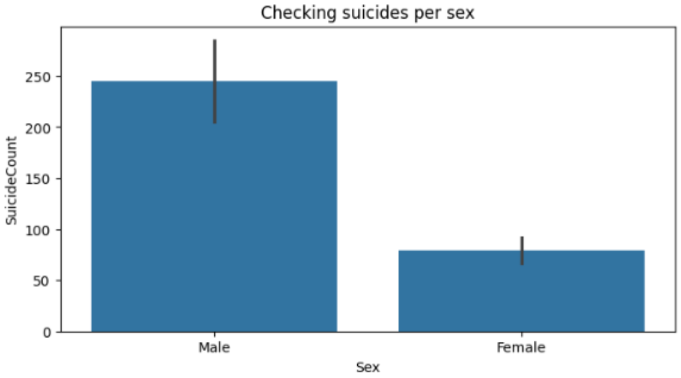
My observations from making a boxplot between sex and the quantity of suicides indicate that men commit suicide more than women.



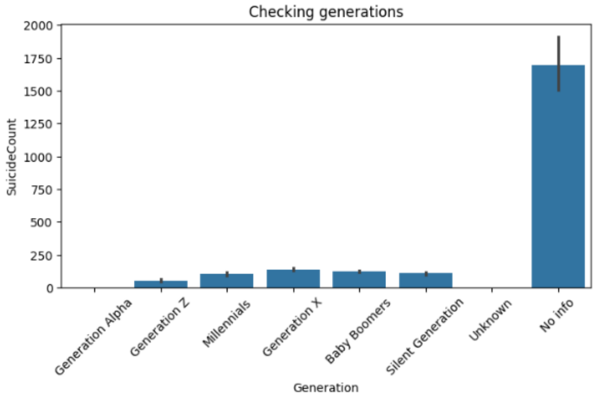
The European suicide boxplot displays more variability in male data than female data, which means that in male, are more different values than in woman. Also tendency higher in men than in woman. Variability is also proportional, pero tampoco se ve muy bien por la gran cantidad de información, es difícil de ver.



The barplot also confirms that the suicide rate in men is higher than in women.



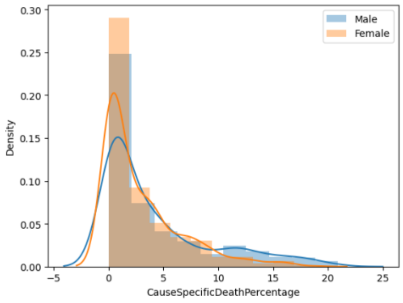
On the other hand, I've checked which generation is most affected by suicide, and that is Generation X, followed by the baby boomers. It's also noticeable that there is a lot of missing information in the dataset regarding this topic

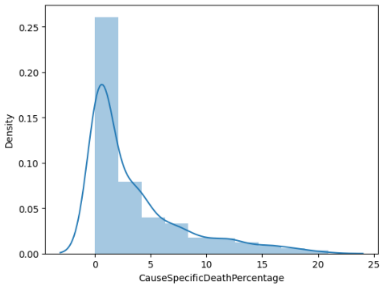


.

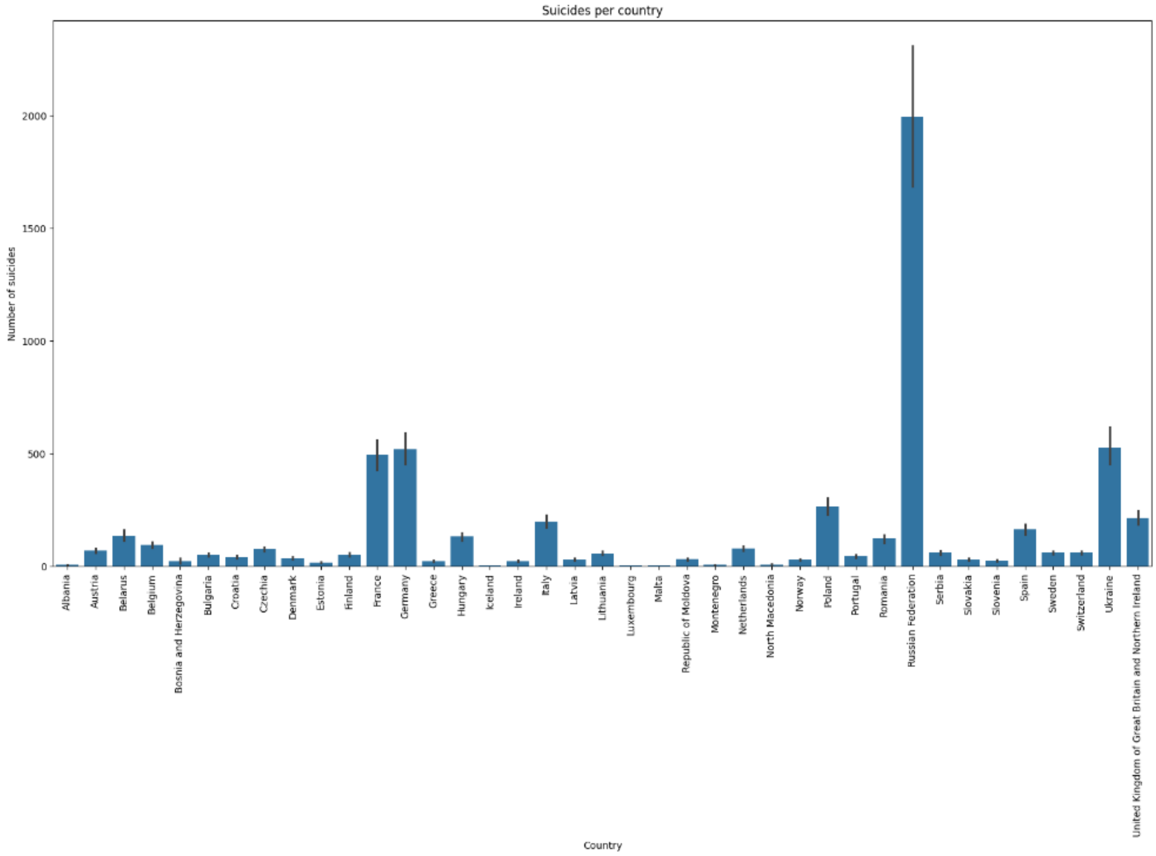
It would have been interesting to analyze the causes of suicide, but even though there is a column for causes, the information in it is purely numerical and does not provide a specific cause. I have tried at least to see the difference by sex, and it can be seen that certain causes affect women much more than men, and others affect men more.

The ones that affect women the most and to a greater extent.The density curve for men has a higher peak than that of women. This indicates a higher density at that specific percentage of cause of death for men compared to women. Overlapping Areas: The areas under each curve are colored: blue for men and orange for women. Some parts of the graph overlap, suggesting shared percentages between the sexes for certain suicide causes.

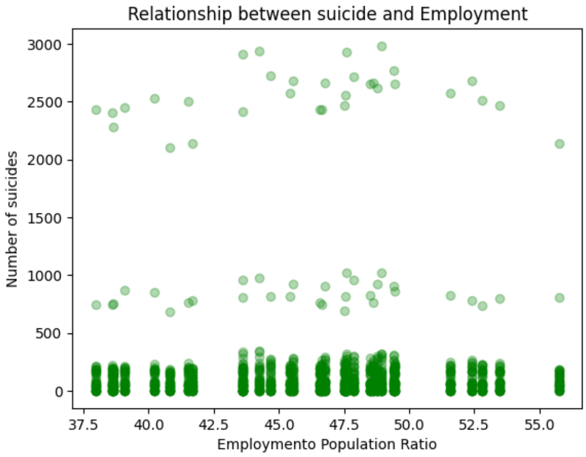




The graph that I show below is the reason why I have created a dataframe with the information from Europe, I wanted to see the number of suicides by European country and see the visual comparison with Spain: I will omit the information from Russia because it is not part of the European Union so first of all we highlight Germany as the leader in suicides and then France and the Ukraine. The situation of Spain with respect to other European countries is moderate or intermediate, but it does not fall within the European countries with the lowest registered suicide rate.



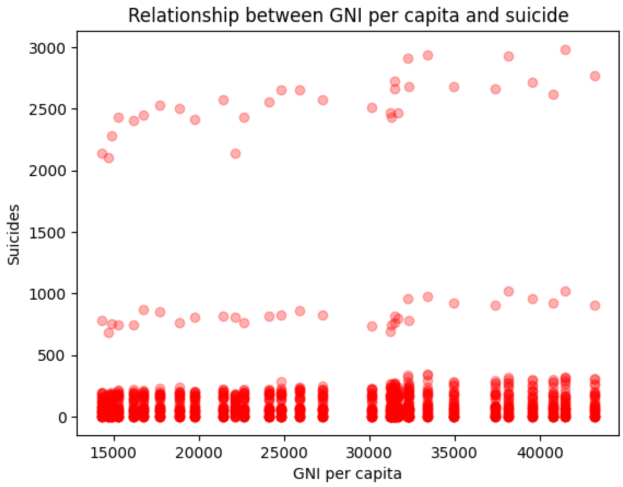
This is a scatterplot that shows a trend in which suicides tend to increase when the employment-population ratio decreases. I think this relationship may be interesting but we will see later what determines the model.



Next scatterplot displays that points are mainly concentrated at the bottom of the vertical axis (number of suicides).

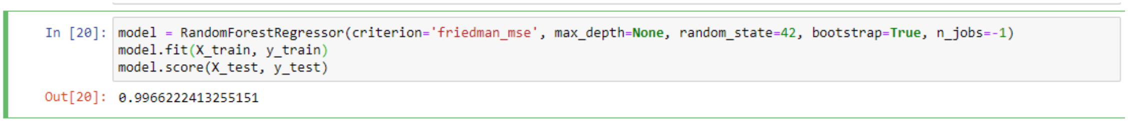
This suggests that there is a higher density of suicides in regions with lower GNP per capita.

In areas with less economic wealth, there could be a higher incidence of suicide.



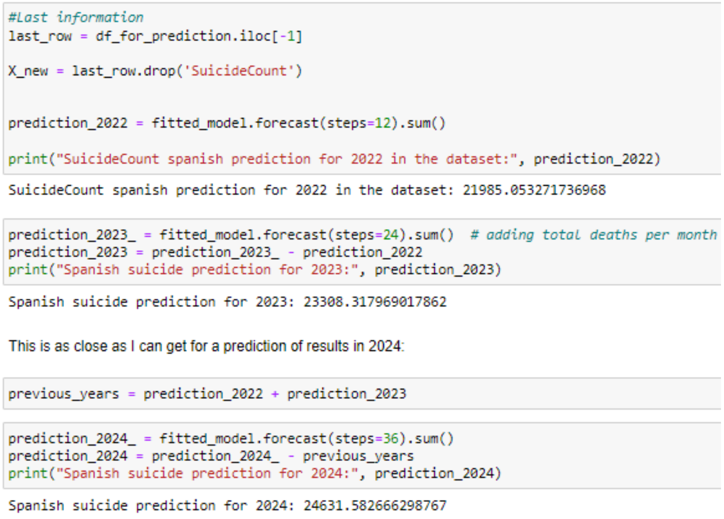
MODEL

Regarding the model to use, I opted for the random forest because there are too many factors to consider for this model and it required discard decisions. Additionally, other models I tried (such as logistic regression model for gender category in relation to suicides, or linear model) did not fit well. Despite consulting with ChatGPT about what was wrong and making modifications, it was not useful; I kept getting a very low coefficient. However, with random forest, I achieved a 0.996...



```
In [20]: model = RandomForestRegressor(criterion='friedman_mse', max_depth=None, random_state=42, bootstrap=True, n_jobs=-1)
         model.fit(X_train, y_train)
         model.score(X_test, y_test)

Out[20]: 0.9966222413255151
```

After building the model, I turned to exponential smoothing, which is a time series forecasting method used to analyze and predict future values based on historical data. In this case, I used it because I checked with value counts that only provide information up to the year 2021. Therefore, I made a prediction for the following year (2022) and also for 2023. Then, I created another code for the current year, which is 2024, and subtracted both values to exclude the entirety. My conclusion is that each year, suicide rates in Spain are projected to increase, which should be a cause for concern.



```
#Last information
last_row = df_for_prediction.iloc[-1]

X_new = last_row.drop('SuicideCount')


prediction_2022 = fitted_model.forecast(steps=12).sum()

print("SuicideCount spanish prediction for 2022 in the dataset:", prediction_2022)

SuicideCount spanish prediction for 2022 in the dataset: 21985.053271736968
```

```
prediction_2023_ = fitted_model.forecast(steps=24).sum()  # adding total deaths per month
prediction_2023 = prediction_2023_ - prediction_2022
print("Spanish suicide prediction for 2023:", prediction_2023)

Spanish suicide prediction for 2023: 23308.317969017862
```

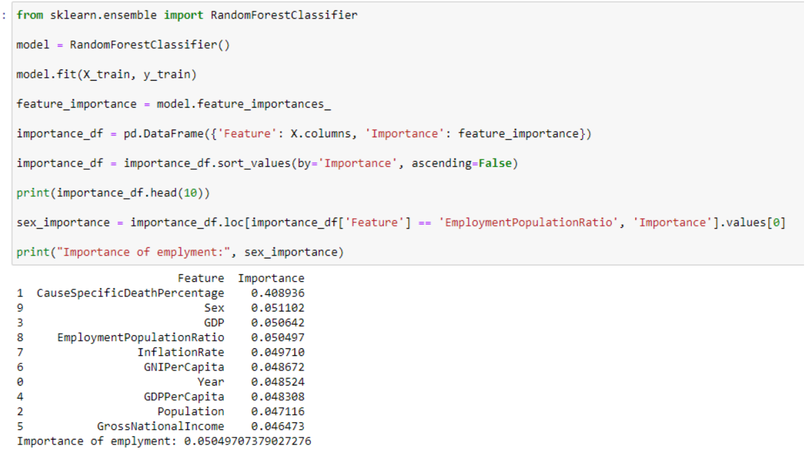
This is as close as I can get for a prediction of results in 2024:

```
previous_years = prediction_2022 + prediction_2023
```

```
prediction_2024_ = fitted_model.forecast(steps=36).sum()
prediction_2024 = prediction_2024_ - previous_years
print("Spanish suicide prediction for 2024:", prediction_2024)

Spanish suicide prediction for 2024: 24631.582666298767
```

I have deemed it appropriate to conduct a significance test on employment as it may provide relevant information about these characteristics and suicide. Surprisingly, despite the previous matplotlib graph indicating a correlation between employment and the suicide rate, which is logical, this model determines that it is not as significant. It assigns it the eighth position among all the features, out of a total of ten (counting from 0 to 9).

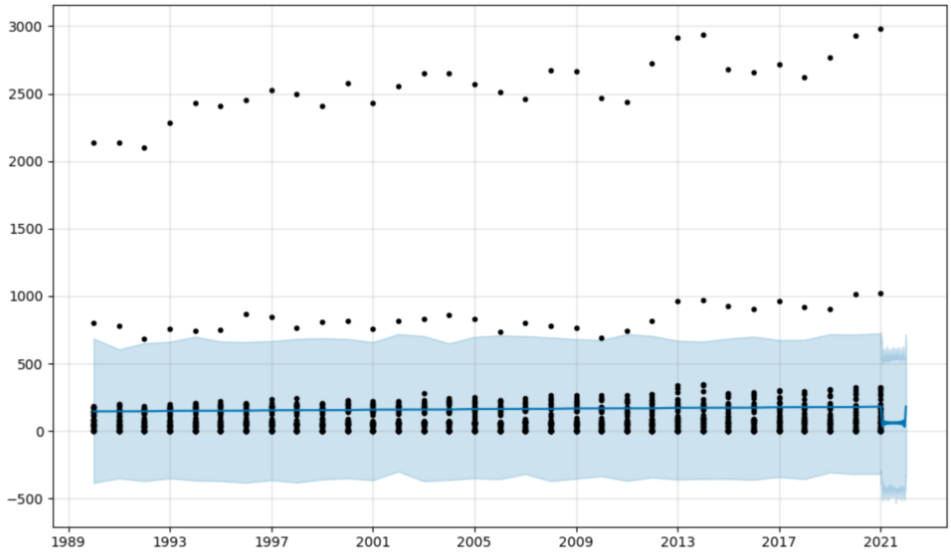
And even though GNI per capita ranks sixth, I believe that the features ahead, although relevant for calculations, are not as crucial for determining reasons in this test. Therefore, I would emphasize that the income of each individual in a country should indeed be considered and relevant. There may be plenty of employment, but not of quality to understand how one's financial situation might influence fatal decisions. Additionally, this dataset addresses specific causes but does not indicate what they are or whether these jobs are of quality. Hence, I conclude that GNI per capita is important and impactful.

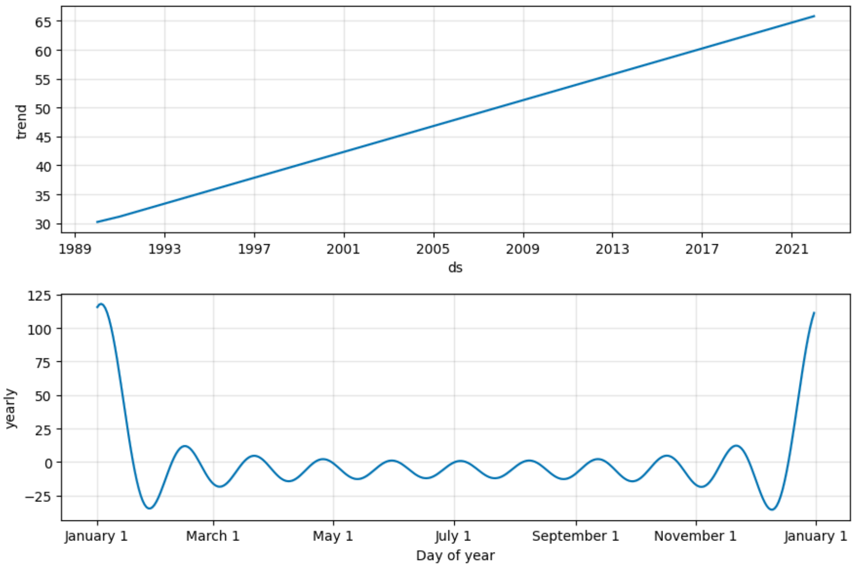


```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier()

model.fit(X_train, y_train)

feature_importance = model.feature_importances_

importance_df = pd.DataFrame({'Feature': X.columns, 'Importance': feature_importance})

importance_df = importance_df.sort_values(by='Importance', ascending=False)

print(importance_df.head(10))

sex_importance = importance_df.loc[importance_df['Feature'] == 'EmploymentPopulationRatio', 'Importance'].values[0]

print("Importance of emplyment:", sex_importance)
```

```
                        Feature  Importance
1  CauseSpecificDeathPercentage    0.408936
9                           Sex    0.051102
3                           GDP    0.050642
8     EmploymentPopulationRatio    0.050497
7                 InflationRate    0.049710
6                  GNIPerCapita    0.048672
0                          Year    0.048524
4                  GDPPerCapita    0.048308
2                    Population    0.047116
5           GrossNationalIncome    0.046473
Importance of emplyment: 0.05049707379027276
```

I have used Prophet to further support the content to be displayed, and as seen in the resulting graphs, it confirms once again that there is a general increase in suicide in Spain and it fluctuates depending on the seasons.




trend
65
60
55
50
45
40
35
30
1989
1993
1997
2001
2005
2009
2013
2017
2021
ds
yearly
125
100
75
50
25
0
−25
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year