Project Report

# GitHub URL

<https://github.com/BorisMoric/UCDPA_BorisMoric>

# Abstract

The aim of the project was to use Python and available read world data and generate valuable insights into question about factors that may lead to people’s decisions on having their kids at different times in their lives. The project used three data sets and was aiming to analyse and manipulate the data contained to enable better understanding of the contents and insights generation.

# Introduction

During November the family was watching Netflix documentary “Pepsi, Where’s My Jet?”. At the same time at UCD-CIDAB course we have just completed 3rd session when it was recommended to start thinking about the project ideas and related data sets. At one point my younger son asked a question “*Do people in rich countries have kids when they are older compared to poorer countries?*”. That sparked a short but lively discussion within the family as to *the reasons that may lead to a decision to have kids & when to have them in ones life*, which gave me the motivation to start looking for data sources that would provide some insights.

# Dataset

## Births registration

The first question to be answered was what public data exists about the births. By searching through the web I concluded that this info was collected and published by the governments of the countries for the country in question. Is there a public data set out there that combines the info from multiple countries. There is an interesting one I found on (UNICEF DATA, 2022). The overall file was 656Mb in size and was locally named “Percent of children\_UNICEF\_1.0\_all”. It contained 1,648,747 rows with 22 columns. The column names and short descriptions for each is contained within .iptnb file.

## Age of mothers at childbirth

From analysis of the birth registrations dataset (UNICEF DATA, 2022) it was concluded that it did not contain the information about the age of parents at the time of the first birth of their child so another source was found from (OECD) that seem to have that. The web site provides a number of datasets. The specific data set used was listed under ‘Fertility indicators’ section SF2.3. Generated file ‘*SF\_2\_3\_Age\_moth.xlsx*’ contained several sheets. Source data from xls was converted into csv as it was easier to work with csv in python. The XLS sheet ‘*Mean-age-first-birth*’ was converted to CSV file ‘*Mean-1st-birth.csv*’.

## Countries of the world

To get some insights into why people have babies at different times in their lives there was a need to add some more parameters into the mix like where in the world the countries are located, ie. are there any regional factors, if there is a correlation with wealth, countries population sizes and etc. “Countries of the world” dataset (Lasso, 2017) contains World fact sheet, with Information on population, region, area size, infant mortality and more. The data set was downloaded by using Keggle API (Kaggle, n.d.).

# Implementation Process

To start learning about “*the reasons that may lead to a decision to have kids & when to have them in ones life*” I started to first look at what data sources are available on the subject of child births. “Births registration” data from UNICEF was the starting point (UNICEF DATA, 2022).

The dataset was imported into dataframe through *pd.read\_csv()* function. Then by using some basic python methods and functions like *df.shape*, *df.info()*, groupby(), *nunique()*it was easy to start getting the feel for the data set which in a hindsight was too big (and too scarry) for my very first python project. After identifying the correct column and exact value to slice the data down (by passing the unique column values through *index.str.contains()*) the data set was reduced quite a bit. Next step was to drop all of the rows that only contained null values. At this stage I decided to reduce the set even lower and decided to eliminate countries that have low % (<80%) of registered births not to influence the findings negatively. Resulting data set was reduced to 75 countries. Countries names were also cleaned up by removing 3 letter acronyms.

Second dataset “Age of mothers at childbirth” (OECD) dataset was also imported with *read\_CSV()* with the intention to combine the two (inner merge on countries). This was much smaller dataset and after replacing “..” characters with NaN’s and removing countries with no data, the dataset resulted in 39 countries and was ready for merge with previous one. Planned merge, however didn’t go as planned, as didn’t yield any overlap, apart from Turkey which was spotted later. This was a surprise and resulted in [1st insight](#_Across_all_of) which was on the importance of and difficulty in selecting the right data sources.

The [2nd insight](#_Across_all_of_1) was based on “Age of mothers at childbirth” (OECD) dataset, see *Figures 1.2.3* and *4*. in the next chapter. The graphs were produced using seaborn package and related functionality.

Now was time to bring in more information from the “Countries of the world” dataset. This dataset was downloaded from the Web by using Kaggle API (Kaggle, n.d.). After using similar techniques to analyse the it I noticed that majority of columns with numerical data (ie. float looking but strings) contained colon instead of the decimal point and wrote a custom function to help replace the characters. At this point I've noticed that there is a space at the end of each country name in CotW\_no\_NaNs dataset. This needed to be removed as merge function would not match the indexes properly. After that merge with “Age of mothers at childbirth” dataset went smoothly and we ended up with 37 countries and 80 columns of data.

Additional insights were derived by simply sorting the resulting dataset by GDP per capita after which it was clear that in poorer countries mothers are having their 1st born earlier when compared with ones from more richer countries [3rd insight](#_There_is_a). All of the wealthiest countries *(Figure 6.)* had significantly higher age of mothers when compared with poorer countries with US being the only outlier in this case with all of the other wealthiest countries coming from EU ([4th insight](#_All_of_the)). Then looking further, from the current EU member states, the countries listed among the top and bottom quartile by wealth all of the poorer ones listed on *Figure 5. w*ere not members prior to year 2004 ([5th insight](#_From_the_current)).

Following similar logic it would have been possible to deduce many more insights (e.g. by looking at the literacy and migration levels or infant mortality). This is as many people do make a lot of life changing decisions – like having a first born – considering their own financial position.

# Results

Charts below are derived from (OECD) report we looked at where 39 countries have reported the yearly average age of the mothers at first child birth in the span of last 60 years.

While merging the (UNICEF DATA, 2022) and (OECD) datasets with >80% condition of reported births it was noted that ([1st insight](#_Across_all_of)) there is no overlap – apart from Turkey – even when relaxing the “>80%” condition and looking at all of the 120 countries from (UNICEF DATA, 2022) which would indicate that either the datasets are incomplete or same data from the countries are not equally collected, analysed and published by both organisations.

Below is copy/paste output from jupyter lab notebook under chapter … >>>

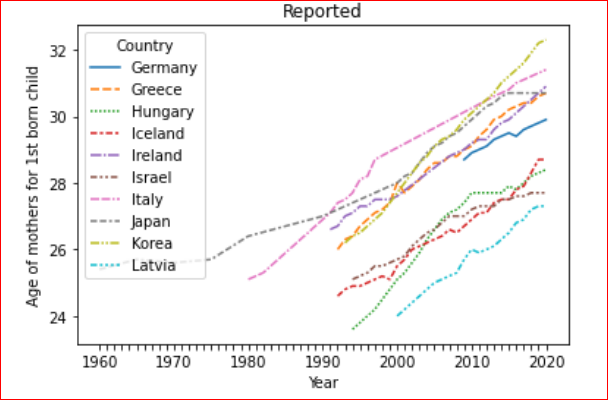
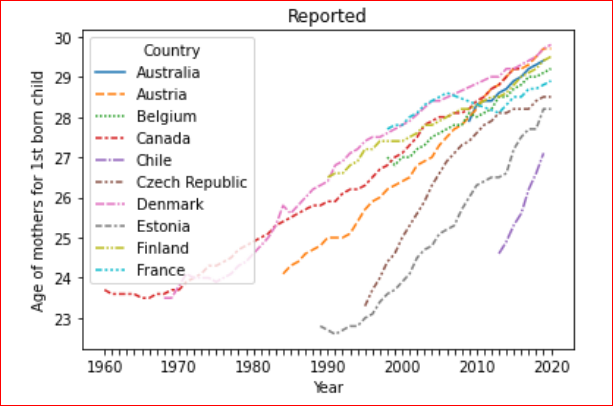
“Births registration” (UNICEF DATA, 2022)

['Afghanistan', 'Albania', 'Algeria', 'Angola', 'Argentina', 'Armenia', 'Azerbaijan', 'Bahrain', 'Bangladesh', 'Barbados', 'Belize', 'Benin', 'Bhutan', 'Botswana', 'Burkina Faso', 'Burundi', 'Cabo Verde', 'Cambodia', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo', 'Cuba', "Côte d'Ivoire", "Democratic People's Republic of Korea", 'Democratic Republic of the Congo', 'Dominican Republic', 'Egypt', 'El Salvador', 'Equatorial Guinea', 'Eswatini', 'Ethiopia', 'Fiji', 'Gabon', 'Gambia', 'Georgia', 'Ghana', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti', 'Honduras', 'India', 'Indonesia', 'Iran (Islamic Republic of)', 'Iraq', 'Jamaica', 'Jordan', 'Kazakhstan', 'Kenya', 'Kiribati', 'Kosovo', 'Kyrgyzstan', "Lao People's Democratic Republic", 'Lebanon', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Maldives', 'Mali', 'Marshall Islands', 'Mauritania', 'Mexico', 'Mongolia', 'Montenegro', 'Morocco', 'Mozambique', 'Myanmar', 'Namibia', 'Nauru', 'Nepal', 'Niger', 'North Macedonia', 'Pakistan', 'Panama', 'Papua New Guinea', 'Paraguay', 'Peru', 'Philippines', 'Republic of Moldova', 'Rwanda', 'Saint Lucia', 'Samoa', 'Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia', 'Sierra Leone', 'Singapore', 'Slovakia', 'Solomon Islands', 'Somalia', 'South Sudan', 'Sri Lanka', 'State of Palestine', 'Sudan', 'Suriname', 'Tajikistan', 'Thailand', 'Timor-Leste', 'Togo', 'Tonga', 'Trinidad and Tobago', 'Tunisia', 'Turkmenistan', 'Turks and Caicos Islands', 'Tuvalu', 'Türkiye', 'Uganda', 'Ukraine', 'United Republic of Tanzania', 'Uruguay', 'Vanuatu', 'Viet Nam', 'Yemen', 'Zambia', 'Zimbabwe']

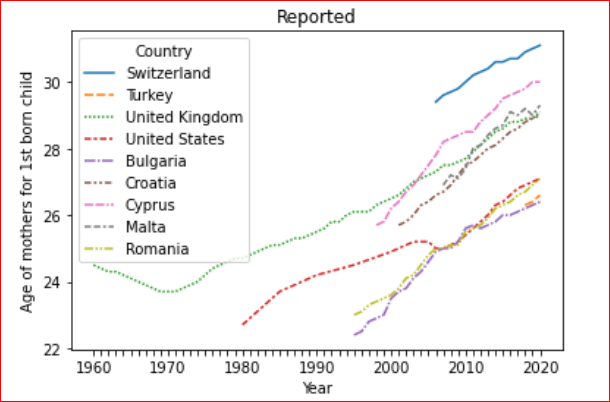
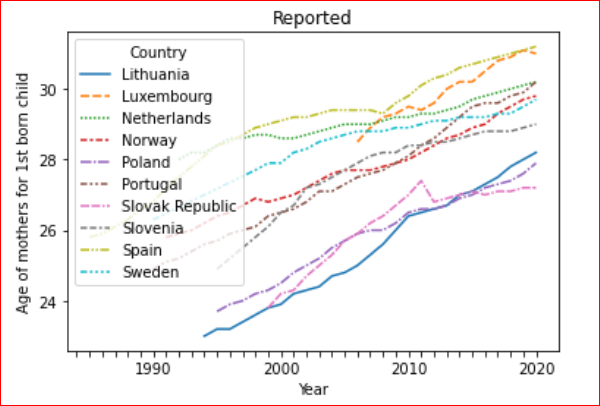
“Age of mothers at childbirth” (OECD)

['Australia', 'Austria', 'Belgium', 'Canada', 'Chile', 'Czech Republic', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Japan', 'Korea', 'Latvia', 'Lithuania', 'Luxembourg', 'Netherlands', 'Norway', 'Poland', 'Portugal', 'Slovak Republic', 'Slovenia', 'Spain', 'Sweden', 'Switzerland', 'Turkey', 'United Kingdom', 'United States', 'Bulgaria', 'Croatia', 'Cyprus', 'Malta', 'Romania']

Our [2nd insight](#_Across_all_of_1) can be derived from observing the trends in firstborns. Across all of the countries during last 60 years we can clearly see that people are deciding to have their first child later in life.



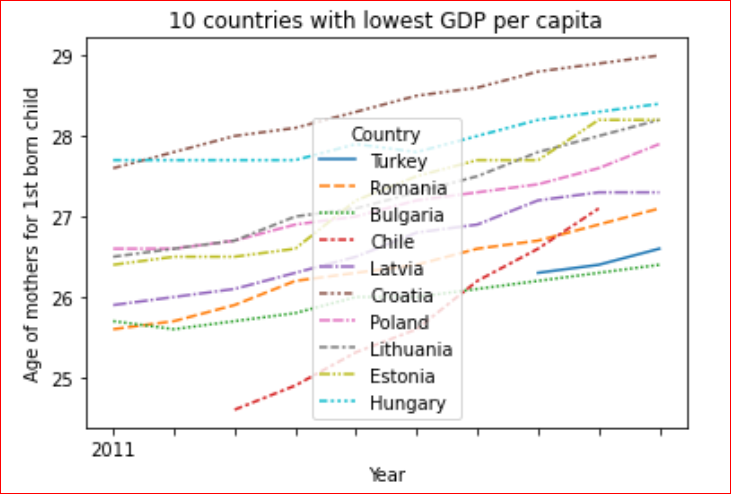
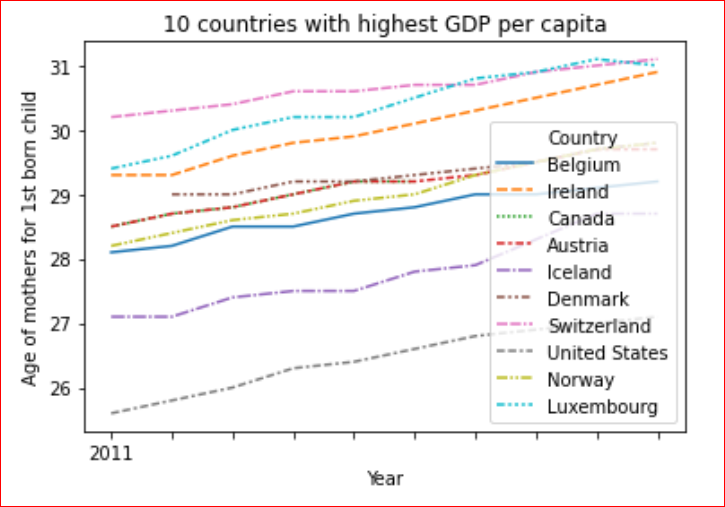
*Figure 1. Figure 2.*



*Figure 3. Figure 4.*

The [3rd insight](#_There_is_a) was derived by using the dataset that combined additional information including GDP per capita. Intuitively, it would have been expected that there is a correlation, and the easy way to show it, was to simply sort the dataset by the GDP per capita column and re-draw the same graphs but in this instance we would have the countries divided in quartiles by GDP.

Additional insights ([4th](#_All_of_the) and [5th](#_From_the_current)) were deduced by using similar observations.

*Figure 5. Figure 6.*

On a reflection, the 1st insight (and a learning point) was the most valuable from the business view point as it pointed out that having a good datasets to work with is crucial in having the tools to make quick (and insightful) business decisions.

On the machine learning topic, it can be observed that many countries haven’t been collecting the data in the past hence data points are missing from the graphs. And if we wanted to use ML in predicting further trends in the future or looking in the past (where the figures were missing) then regression type of analysis would be most suited. This is because the data is continuous by nature and there is no simple (yes/no) factors (ie. Classification) that would play a major role while deciding to have children.

# Insights

*(Point out at least 5 insights in bullet points)*

## There was no overlap between countries with reported childbirths from (UNICEF DATA, 2022) dataset and Age of mothers at 1st child birth from (OECD) dataset (apart from Turkey).

## Across all of the countries during last 60 years of reported data there is a clear trend of having a firstborn later and later in life.

## 

## There is a correlation between the country’s wealth (indicator used was GDP per capita) and the time mothers in those geographies are having their 1st born babies.

## 

## All of the wealthiest countries had significantly higher age of mothers when compared with poorer countries with US being the only outlier in this case with all of the other wealthiest countries coming from EU

## 

## From the current EU member countries listed among the top and bottom countries by wealth all of the ones listed on Figure 5. were not members prior to year 2004

# References

Kaggle. (n.d.). *Public API documentation | Kaggle*. Retrieved from https://www.kaggle.com/docs/api

Lasso, F. (2017). *Countries of the World | Kaggle.* Retrieved 12 30, 2022, from kaggle.com: https://www.kaggle.com/datasets/fernandol/countries-of-the-world

OECD. (n.d.). *OECD Family Database - OECD.* Retrieved Dec 30, 2022, from oecd.org: https://www.oecd.org/els/family/database.htm (Excel file was accessible through: https://www.oecd.org/els/soc/SF\_2\_3\_Age\_mothers\_childbirth.xlsx)

UNICEF DATA. (2022, May). *Birth registration data - UNICEF DATA data.unicef.org.* Retrieved Dec 30, 2022, from data.unicef.org.: https://data.unicef.org/resources/dataset/percentage-children-age-5-whose-births-registered-sex-place-residence-household-wealth-quintile/