Children per woman (total fertility rate)

Table of Contents

Introduction

Children per woman (total fertility rate) all over the world.

Sources

— 1800 to 1950 (and in some cases also years after 1950): Gapminder v6 which were compiled and documented by Mattias Lindgren

— 1950 to 2014: In most cases we use the latest UN estimates from World Population Prospects 2017 published in the file with Annually interpolated demographic indicators, called WPP2017_INT_F01_ANNUAL_DEMOGRAPHIC_INDICATORS.xlsx , accessed on September 2, 2017.

— 2015 – 2099: We use the UN forecast of future fertility rate in all countries, called median fertility variant.

• For more information about how the sources were combined and download link please visit this link:

https://www.gapminder.org/data/documentation/gd008/

The main questions that came to my mind when I first known about this study:

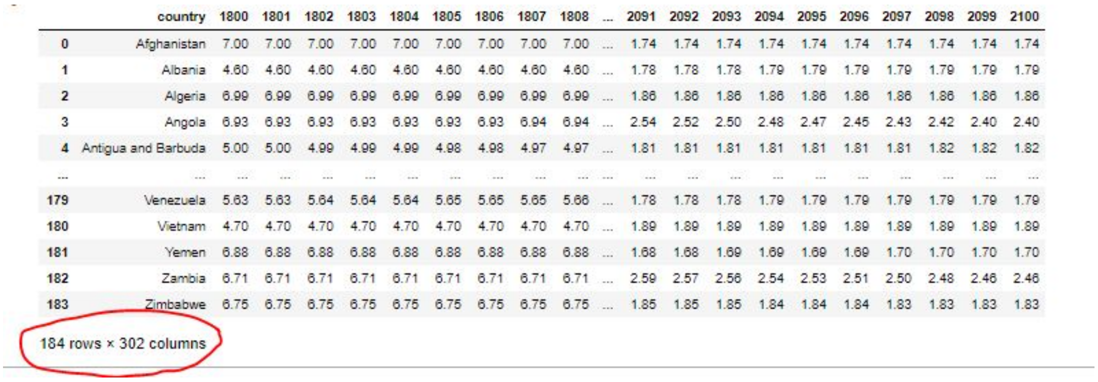
1) Is there a certain fertility rate trend throughout the time?
2) Does the rate get influenced by the location on earth?
3) Is there a specific nation that maintained the highest fertility rate or it changed over time?

In this analysis I will be trying to find answers to these questions so follow along :)

Data Wrangling

I will summarize my steps to make it easy to follow and for the code source I will attached it with the files I'll upload in submission

1) The data set downloaded from this link was like this:



	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	...	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100
0	Afghanistan	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	...	1.74	1.74	1.74	1.74	1.74	1.74	1.74	1.74	1.74	1.74
1	Albania	4.60	4.60	4.60	4.60	4.60	4.60	4.60	4.60	4.60	...	1.78	1.78	1.78	1.79	1.79	1.79	1.79	1.79	1.79	1.79
2	Algeria	6.99	6.99	6.99	6.99	6.99	6.99	6.99	6.99	6.99	...	1.86	1.86	1.86	1.86	1.86	1.86	1.86	1.86	1.86	1.86
3	Angola	6.93	6.93	6.93	6.93	6.93	6.93	6.93	6.94	6.94	...	2.54	2.52	2.50	2.48	2.47	2.45	2.43	2.42	2.40	2.40
4	Antigua and Barbuda	5.00	5.00	4.99	4.99	4.99	4.98	4.98	4.97	4.97	...	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.82	1.82	1.82
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
179	Venezuela	5.63	5.63	5.64	5.64	5.64	5.65	5.65	5.65	5.66	...	1.78	1.78	1.78	1.79	1.79	1.79	1.79	1.79	1.79	1.79
180	Vietnam	4.70	4.70	4.70	4.70	4.70	4.70	4.70	4.70	4.70	...	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
181	Yemen	6.88	6.88	6.88	6.88	6.88	6.88	6.88	6.88	6.88	...	1.68	1.68	1.69	1.69	1.69	1.69	1.70	1.70	1.70	1.70
182	Zambia	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	6.71	...	2.59	2.57	2.56	2.54	2.53	2.51	2.50	2.48	2.46	2.46
183	Zimbabwe	6.75	6.75	6.75	6.75	6.75	6.75	6.75	6.75	6.75	...	1.85	1.85	1.85	1.84	1.84	1.84	1.83	1.83	1.83	1.83

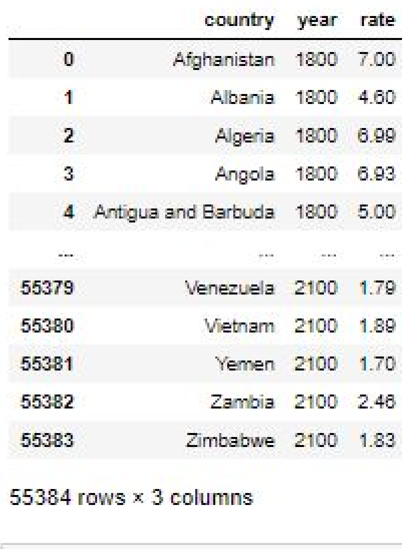
184 rows × 302 columns

Very wide dataframe and untidy structure:

- Years was distributed in columns
- Very wide range of countries (actually the whole world countries 😂) which is very large data to deal

2) cleaning:

- For the 1st issue, I've restructured the data frame sowe have a tidy form to deal in analyzing:

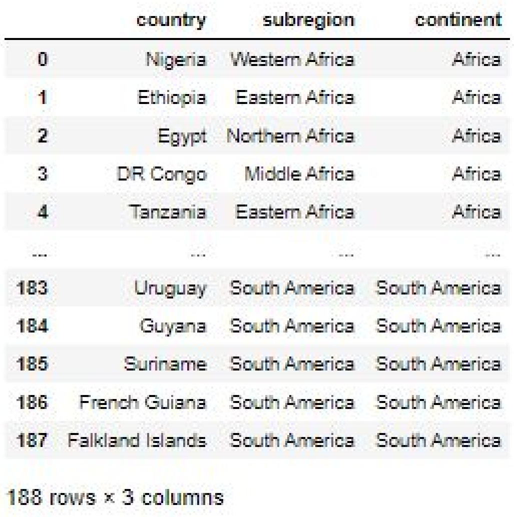


	country	year	rate
0	Afghanistan	1800	7.00
1	Albania	1800	4.60
2	Algeria	1800	6.99
3	Angola	1800	6.93
4	Antigua and Barbuda	1800	5.00
...	...	...	...
55379	Venezuela	2100	1.79
55380	Vietnam	2100	1.89
55381	Yemen	2100	1.70
55382	Zambia	2100	2.46
55383	Zimbabwe	2100	1.83

55384 rows × 3 columns

- for the 2nd issue, I got use the help of this website https://www.worldometers.info/ to categorize the countries into Continents so make a chunks that can be observable.

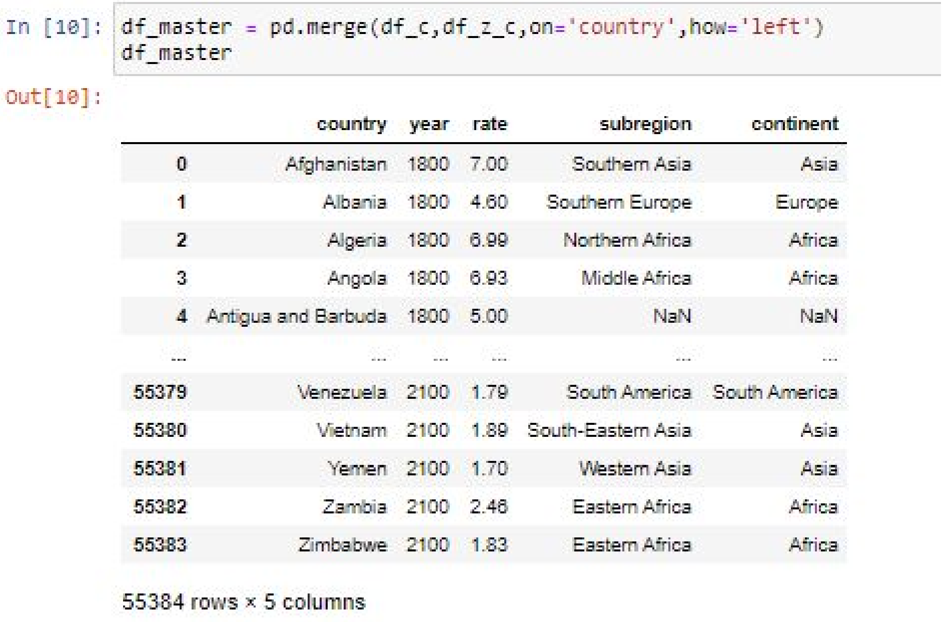
To make this happen I've invented a new data frame with categorization to the world countries



	country	subregion	continent
0	Nigeria	Western Africa	Africa
1	Ethiopia	Eastern Africa	Africa
2	Egypt	Northern Africa	Africa
3	DR Congo	Middle Africa	Africa
4	Tanzania	Eastern Africa	Africa
...	...	...	...
183	Uruguay	South America	South America
184	Guyana	South America	South America
185	Suriname	South America	South America
186	French Guiana	South America	South America
187	Falkland Islands	South America	South America

188 rows × 3 columns

3) **Then I created the master data set with simi-tidy form and enough information to begin**



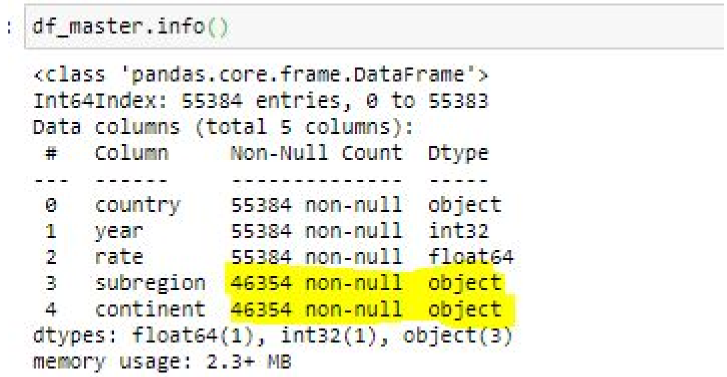
```
In [10]: df_master = pd.merge(df_c,df_z_c,on='country',how='left')
         df_master
```

Out[10]:

	country	year	rate	subregion	continent
0	Afghanistan	1800	7.00	Southern Asia	Asia
1	Albania	1800	4.60	Southern Europe	Europe
2	Algeria	1800	6.99	Northern Africa	Africa
3	Angola	1800	6.93	Middle Africa	Africa
4	Antigua and Barbuda	1800	5.00	NaN	NaN
...	...	...	...	...	...
55379	Venezuela	2100	1.79	South America	South America
55380	Vietnam	2100	1.89	South-Eastern Asia	Asia
55381	Yemen	2100	1.70	Western Asia	Asia
55382	Zambia	2100	2.46	Eastern Africa	Africa
55383	Zimbabwe	2100	1.83	Eastern Africa	Africa

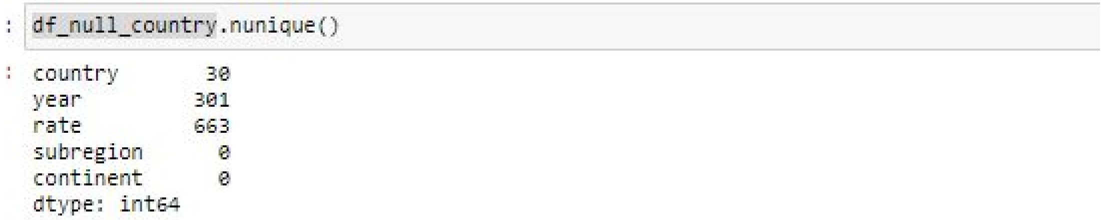
55384 rows × 5 columns

4) **I dealt with this dataframe like it’s a new thing and began to asses t again to chech if it’s ready to make the main dish**
    - **There are missing values for countries categorization**

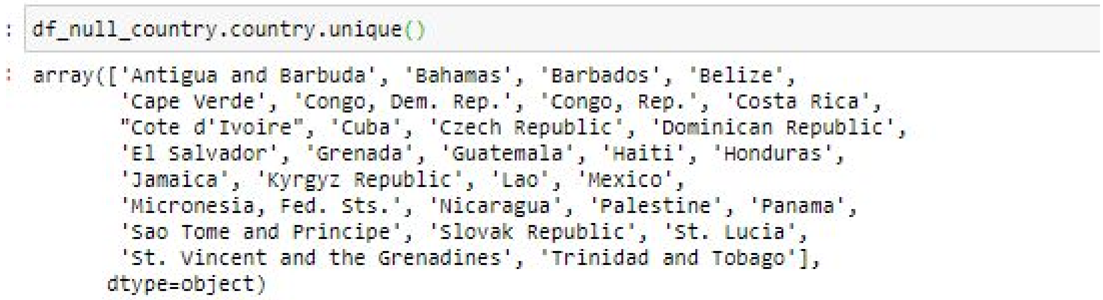


```
: df_master.info()

  <class 'pandas.core.frame.DataFrame'>
  Int64Index: 55384 entries, 0 to 55383
  Data columns (total 5 columns):
   #   Column     Non-Null Count  Dtype  
  ---  ------     --------------  -----  
   0   country    55384 non-null  object 
   1   year       55384 non-null  int32  
   2   rate       55384 non-null  float64
   3   subregion  46354 non-null  object 
   4   continent  46354 non-null  object 
  dtypes: float64(1), int32(1), object(3)
  memory usage: 2.3+ MB
```

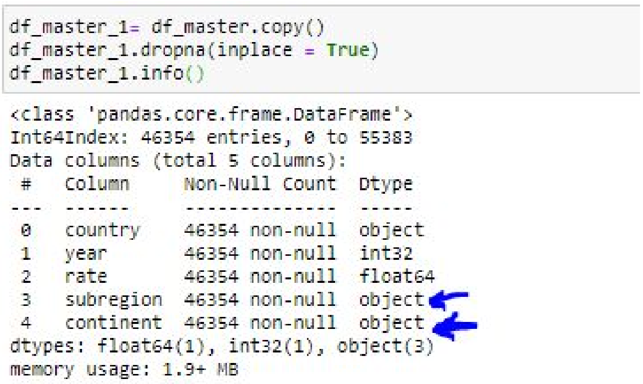
5) **I’ve explored the countries with no categorization and found out they are 30 country out of 184 so it will not make that much difference if they got neglected so I’ve dropped them**



```
: df_null_country.nunique()

: country       30
  year         301
  rate         663
  subregion      0
  continent      0
  dtype: int64
```

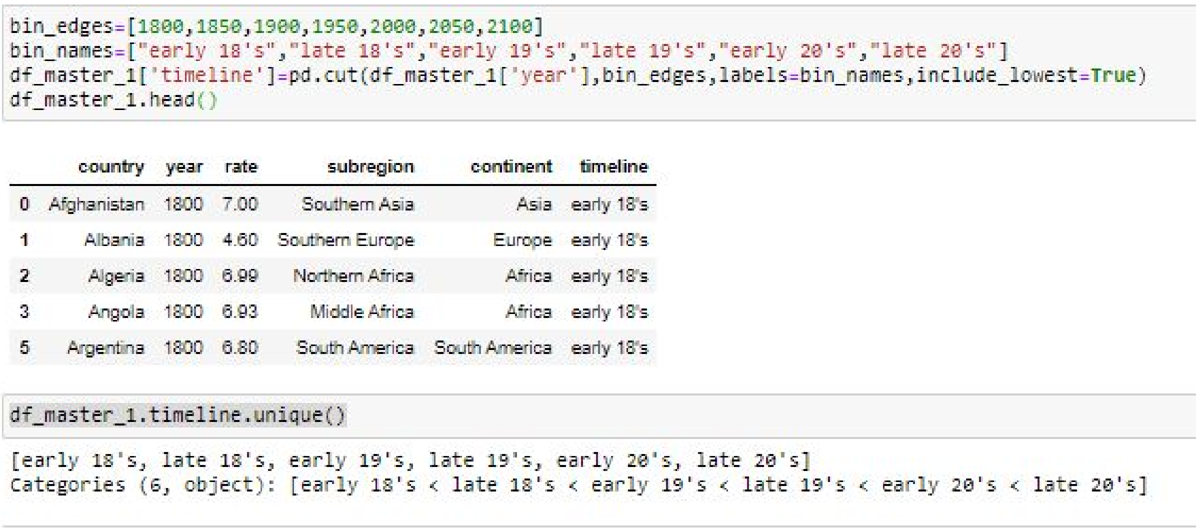


```
: df_null_country.country.unique()

: array(['Antigua and Barbuda', 'Bahamas', 'Barbados', 'Belize',
         'Cape Verde', 'Congo, Dem. Rep.', 'Congo, Rep.', 'Costa Rica',
         "Cote d'Ivoire", 'Cuba', 'Czech Republic', 'Dominican Republic',
         'El Salvador', 'Grenada', 'Guatemala', 'Haiti', 'Honduras',
         'Jamaica', 'Kyrgyz Republic', 'Lao', 'Mexico',
         'Micronesia, Fed. Sts.', 'Nicaragua', 'Palestine', 'Panama',
         'Sao Tome and Principe', 'Slovak Republic', 'St. Lucia',
         'St. Vincent and the Grenadines', 'Trinidad and Tobago'],
        dtype=object)
```



```
df_master_1= df_master.copy()
df_master_1.dropna(inplace = True)
df_master_1.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 46354 entries, 0 to 55383
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   country    46354 non-null  object 
 1   year       46354 non-null  int32  
 2   rate       46354 non-null  float64
 3   subregion  46354 non-null  object 
 4   continent  46354 non-null  object 
dtypes: float64(1), int32(1), object(3)
memory usage: 1.9+ MB
```

6) **Then when I began to explore the data I found out that there is a 301 year record (1800 to 2100) for 184 country so I decided to concat the timeline and created 6 time zone (early 18’s , late 18’s , early 19’s, late 19’s, early 20’s, late 20’s) chuncks of 50 years.**



```
bin_edges=[1800,1850,1900,1950,2000,2050,2100]
bin_names=["early 18's","late 18's","early 19's","late 19's","early 20's","late 20's"]
df_master_1['timeline']=pd.cut(df_master_1['year'],bin_edges,labels=bin_names,include_lowest=True)
df_master_1.head()
```

	country	year	rate	subregion	continent	timeline
0	Afghanistan	1800	7.00	Southern Asia	Asia	early 18's
1	Albania	1800	4.60	Southern Europe	Europe	early 18's
2	Algeria	1800	6.99	Northern Africa	Africa	early 18's
3	Angola	1800	6.93	Middle Africa	Africa	early 18's
5	Argentina	1800	6.80	South America	South America	early 18's

```
df_master_1.timeline.unique()
```

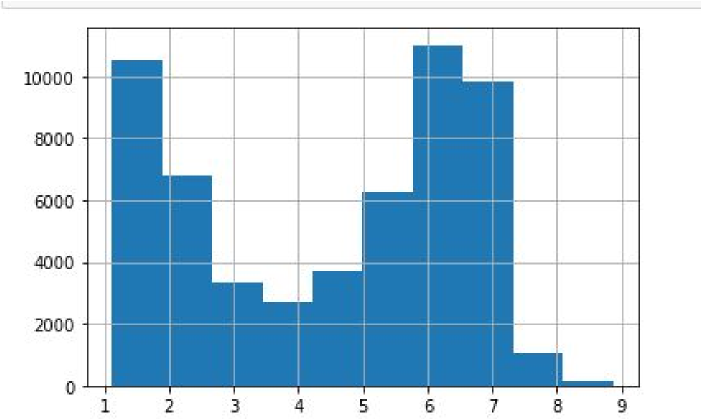
```
[early 18's, late 18's, early 19's, late 19's, early 20's, late 20's]
Categories (6, object): [early 18's < late 18's < early 19's < late 19's < early 20's < late 20's]
```

And now I believe it is ready to dive in :)

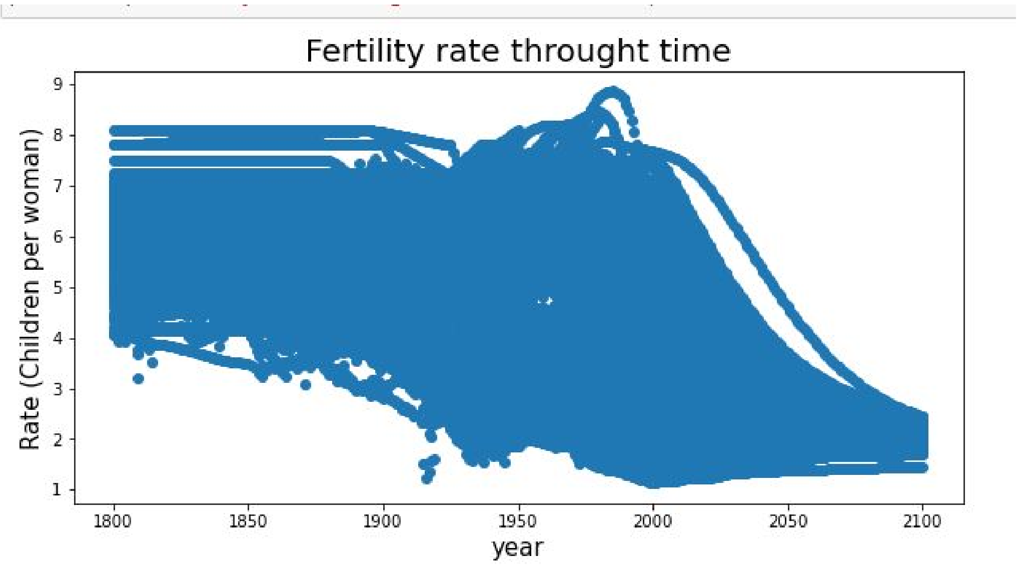
Exploratory Data Analysis:

Once the data is clear and tidy it’s time to dive in the data to get much knowing about it.

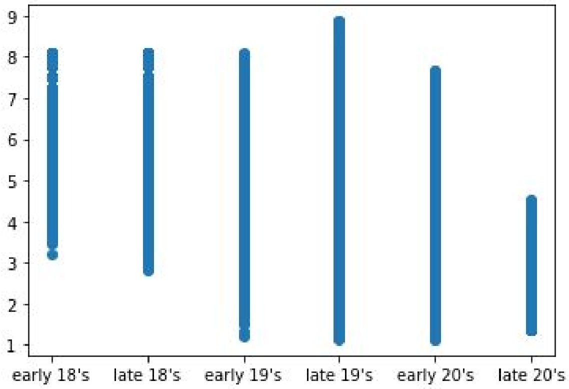
1. **Fertility rates ranges between 1 to 8 child per woman:**



2. **Fertility rate tends to decrease along the time:**



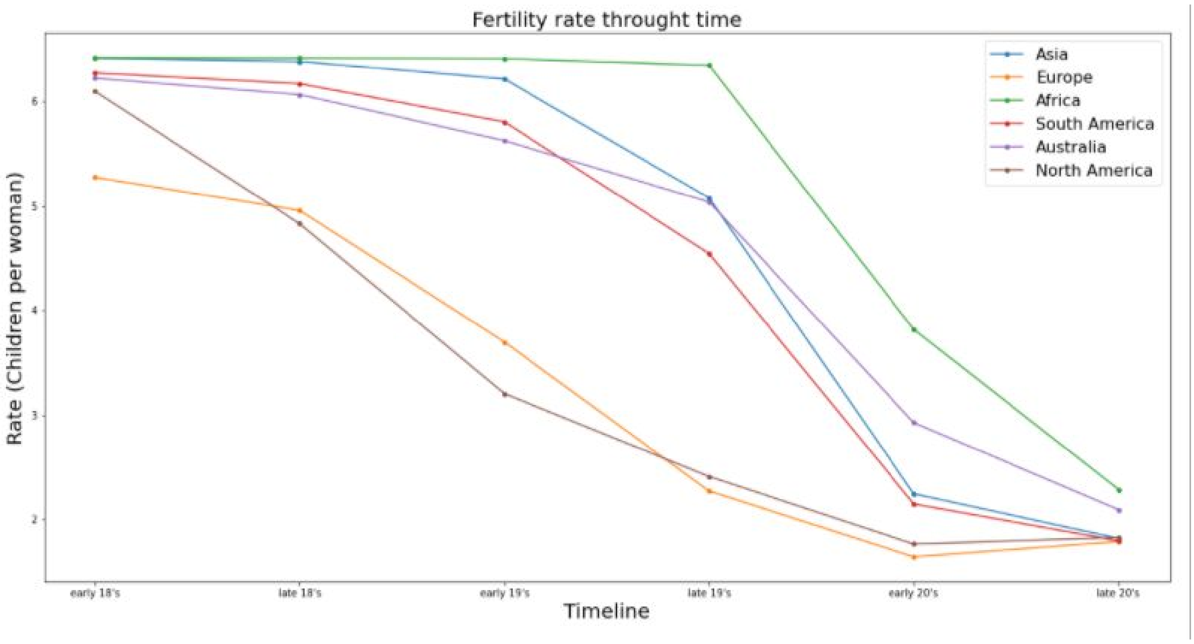
3. **in depth investigations we find fertility rates began to decrease at the early 19’s**



4. The main reason I’ve chosen this type of chart generate is because I think it gives pretty much everything about the data as there are three main variables (place, time, rate)

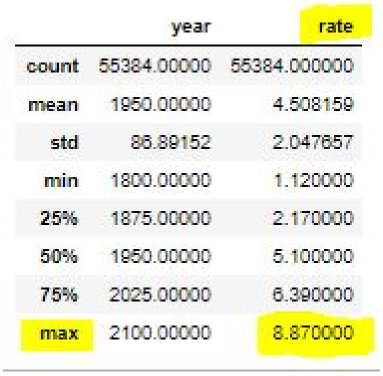
So that I’ve created this graph even it took much effort

Through some coding and organizing, I made a graph that give wide view about the data set



My findings are:

- It's obvious throughout time that mean fertility rate has decreased more and more
- There is no group of people around the world till the meantime (early 20’s) experienced an increase in mean fertility rate but there are expectations for North America and Europe to get higher rates in the future according to the UN forecast of future fertility rate
- Europe and North America experienced an early fertility rate decreasing un like the rest of the world that kind of maintained theri rate and began to decrease roughly on the late 19’s
- The highest fertility rate ever recorded was 8 child per woman



	year	rate
count	55384.00000	55384.000000
mean	1950.00000	4.508159
std	86.89152	2.047657
min	1800.00000	1.120000
25%	1875.00000	2.170000
50%	1950.00000	5.100000
75%	2025.00000	6.390000
max	2100.00000	8.870000

Conclusions

- Limitations to this study:
  - **30 countries didn't get categorized and included in my master data set so it might slightly affect the concat graph of continents and time**
  - **I know from media that there are cases where a mother had over 24 children but it's not included in this database so I had to deal with what in my hand**
- Regarding the questions I've asked earlier here is my conclusion:
  1) **Is there a certain fertility rate trend throughout the time?**

     Yes and it is decreasing throughout time

  2) **Does the rate get influenced by the location on earth?**

     Yes some places have higher mean fertility rate than other like (Asia, Africa, South America, Australia)

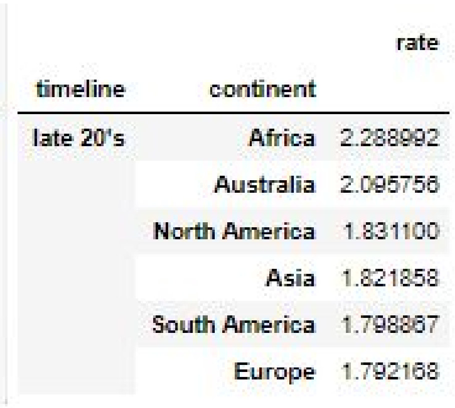
  3) **Is there a specific nation that maintained the highest fertility rate or it changed over time?**

     It is Africa

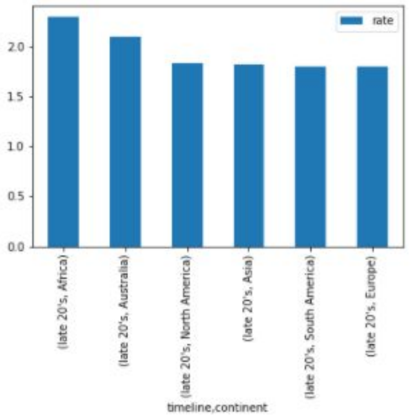
     According to the figure driven from the data. Even in the time of least fertility rate, African countries have the highest mean fertility rates among its peers

- Other notes:

  According to the expected mean fertility rate in the late 20's, the higher rates will be in Africa and Australia (~2 child per woman)



		rate
timeline	continent	
late 20's	Africa	2.288992
	Australia	2.095756
	North America	1.831100
	Asia	1.821858
	South America	1.798867
	Europe	1.792168


rate
(late 20's, Africa)
(late 20's, Australia)
(late 20's, North America)
(late 20's, Asia)
(late 20's, South America)
(late 20's, Europe)
timeline,continent