## Name : Yoginder Singh

## Project : Demographic Dynamics and Health Indicators: A Comprehensive Analysis of Census Data"

## Submission Date :

**The dataset comprises diverse demographic and health indicators, including birth and death rates, fertility, migration, and life expectancy trends across various countries. The process of managing this data involves a sophisticated Data pipeline, beginning with the extraction of specific data points using SQL queries within Google BigQuery. This step ensures efficient and relevant data retrieval from a larger database. Following extraction, the dataset is imported into Google Colab, a cloud-based environment, for exploratory data analysis and cleaning. This stage is crucial for maintaining data integrity and involves addressing missing values, outliers, and potential inaccuracies. The cleaned and analyzed data is then exported to Power BI, where advanced visualization techniques are utilized to create a professional and comprehensive dashboard. This dashboard effectively illustrates key insights through interactive visual elements, making complex demographic patterns and health trends accessible and understandable. The integration of BigQuery, Google Colab, and Power BI in this pipeline underscores a robust approach to data-driven decision-making and insightful analysis..**

```
import matplotlib.pyplot as plt
import seaborn as sns
from google.cloud import bigquery
from google.oauth2 import service_account
import pandas as pd
```

Overview: The United States Census Bureau's international dataset provides estimates of country populations since 1950 and projections through 2050. Specifically, the dataset includes midyear population figures broken down by age and gender assignment at birth. Additionally, time-series data is provided for attributes including fertility rates, birth rates, death rates, and migration rates.

Update frequency: Historic (none)

Dataset source: United States Census Bureau

Terms of use: This dataset is publicly available for anyone to use under the following terms provided by the Dataset Source - http://www.data.gov/privacy-policy#data_policy - and is provided "AS IS" without any warranty, express or implied, from Google. Google disclaims all liability for any damages, direct or indirect, resulting from the use of the dataset.

See the GCP Marketplace listing for more details and sample queries: https://console.cloud.google.com/marketplace/details/united-states-census-bureau/international-census-data

**## Using Big Query to Import data from Google Cloud Platfrom using API Key and SQL ##**

```
# JSON key file
credentials = service_account.Credentials.from_service_account_file(
    '/content/drive/MyDrive/Colab Notebooks/Project/admn5015-winter24-ic-412022-3c9e4269182d.json')

client = bigquery.Client(credentials=credentials, project=credentials.project_id)
```

```
query = """

SELECT
A.country_code,
A.country_name,
A.year,
A.crude_birth_rate,
A.crude_death_rate,
A.net_migration,
A.rate_natural_increase,
A.growth_rate,
B.fertility_rate_15_19,
B.fertility_rate_20_24,
B.fertility_rate_25_29,
B.fertility_rate_30_34,
B.fertility_rate_35_39,
B.fertility_rate_40_44,
B.fertility_rate_45_49,
B.total_fertility_rate,
B.gross_reproduction_rate,
B.sex_ratio_at_birth,
C.infant_mortality,
C.infant_mortality_male,
C.infant_mortality_female,
C.life_expectancy,
C.life_expectancy_male,
C.life_expectancy_female,
C.mortality_rate_under5,
C.mortality_rate_under5_male,
C.mortality_rate_under5_female,
C.mortality_rate_1to4,
C.mortality_rate_1to4_male,
C.mortality_rate_1to4_female
FROM `bigquery-public-data.census_bureau_international.birth_death_growth_rates` as A
JOIN `bigquery-public-data.census_bureau_international.age_specific_fertility_rates` as B
ON
A.country_code = B.country_code
AND A.country_name = B.country_name
AND A.year = B.year
JOIN `bigquery-public-data.census_bureau_international.mortality_life_expectancy` as C
ON
A.country_code = C.country_code
AND A.country_name = C.country_name
AND A.year = C.year;
"""
df = client.query(query).to_dataframe()
```

```
 # first few rows of the dataframe
 df.head()
```

| | country_code | country_name | year | crude_birth_rate | crude_death_rate | net_migration | rate_natural_increase | growth_rate | fertility_ra |
|---|---|---|---|---|---|---|---|---|---|
| **0** | CD | Chad | 1993 | 47.80 | 18.71 | -23.09 | 2.909 | 0.600 | |
| **1** | CD | Chad | 1994 | 47.71 | 18.31 | -0.20 | 2.940 | 2.920 | |
| **2** | CD | Chad | 1995 | 48.02 | 17.90 | 21.95 | 3.012 | 5.207 | |
| **3** | CD | Chad | 1996 | 48.35 | 17.51 | -0.14 | 3.084 | 3.070 | |
| **4** | CD | Chad | 1997 | 48.74 | 17.25 | 0.18 | 3.149 | 3.168 | |

5 rows × 30 columns

```
df.shape
```

```
(15016, 30)
```

```
#Summary of the dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15016 entries, 0 to 15015
Data columns (total 30 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
```

```
 0   country_code                 15016 non-null  object
 1   country_name                 15016 non-null  object
 2   year                         15016 non-null  Int64
 3   crude_birth_rate             15016 non-null  float64
 4   crude_death_rate             15016 non-null  float64
 5   net_migration                15016 non-null  float64
 6   rate_natural_increase        15016 non-null  float64
 7   growth_rate                  15016 non-null  float64
 8   fertility_rate_15_19         15016 non-null  float64
 9   fertility_rate_20_24         15016 non-null  float64
 10  fertility_rate_25_29         15016 non-null  float64
 11  fertility_rate_30_34         15016 non-null  float64
 12  fertility_rate_35_39         15016 non-null  float64
 13  fertility_rate_40_44         15016 non-null  float64
 14  fertility_rate_45_49         15016 non-null  float64
 15  total_fertility_rate         15016 non-null  float64
 16  gross_reproduction_rate      15016 non-null  float64
 17  sex_ratio_at_birth           15016 non-null  float64
 18  infant_mortality             15016 non-null  float64
 19  infant_mortality_male        15016 non-null  float64
 20  infant_mortality_female      15016 non-null  float64
 21  life_expectancy              15016 non-null  float64
 22  life_expectancy_male         15016 non-null  float64
 23  life_expectancy_female       15016 non-null  float64
 24  mortality_rate_under5        15016 non-null  float64
 25  mortality_rate_under5_male   15016 non-null  float64
 26  mortality_rate_under5_female 15016 non-null  float64
 27  mortality_rate_1to4          15016 non-null  float64
 28  mortality_rate_1to4_male     15016 non-null  float64
 29  mortality_rate_1to4_female   15016 non-null  float64
dtypes: Int64(1), float64(27), object(2)
memory usage: 3.5+ MB
```

```
# Checking for the missing value inthe data set
df.isnull().sum()
```

```
country_code                 0
country_name                 0
year                         0
crude_birth_rate             0
crude_death_rate             0
net_migration                0
rate_natural_increase        0
growth_rate                  0
fertility_rate_15_19         0
fertility_rate_20_24         0
fertility_rate_25_29         0
fertility_rate_30_34         0
fertility_rate_35_39         0
fertility_rate_40_44         0
fertility_rate_45_49         0
total_fertility_rate         0
gross_reproduction_rate      0
sex_ratio_at_birth           0
infant_mortality             0
infant_mortality_male        0
infant_mortality_female      0
life_expectancy              0
life_expectancy_male         0
life_expectancy_female       0
mortality_rate_under5        0
mortality_rate_under5_male   0
mortality_rate_under5_female 0
mortality_rate_1to4          0
mortality_rate_1to4_male     0
mortality_rate_1to4_female   0
dtype: int64
```

**Exploratory Data Analysis (EDA)**

```
df1 = df.drop('year', axis = 1)
df1.describe()
```

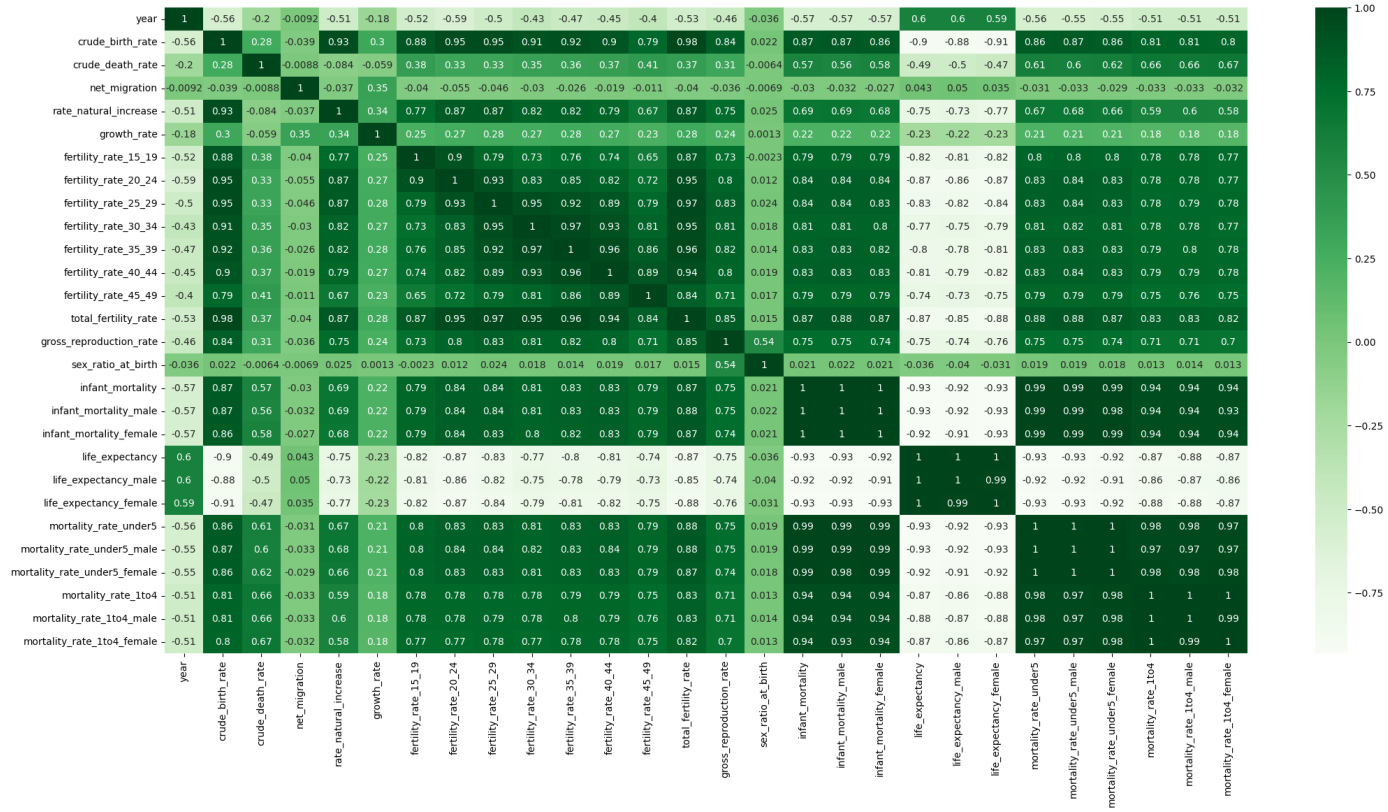
| | crude_birth_rate | crude_death_rate | net_migration | rate_natural_increase | growth_rate | fertility_rate_15_19 | fertility_rate_20_2 |
|---|---|---|---|---|---|---|---|
| **count** | 15016.000000 | 15016.000000 | 15016.000000 | 15016.000000 | 15016.000000 | 15016.000000 | 15016.00000 |
| **mean** | 20.605202 | 8.961052 | -0.229858 | 1.164401 | 1.103004 | 46.048815 | 122.89374 |
| **std** | 11.705203 | 4.188454 | 29.947897 | 1.128683 | 3.189014 | 45.203696 | 73.57341 |
| **min** | 3.620000 | 1.170000 | -831.810000 | -3.265000 | -168.944000 | 0.000000 | 9.50000 |
| **25%** | 11.147500 | 6.170000 | -2.620000 | 0.242000 | 0.200750 | 13.400000 | 66.80000 |
| **50%** | 16.600000 | 8.010000 | -0.160000 | 1.029000 | 0.921000 | 27.900000 | 95.30000 |
| **75%** | 27.912500 | 10.800000 | 1.380000 | 2.099000 | 2.006250 | 63.900000 | 166.90000 |
| **max** | 58.740000 | 65.430000 | 1693.010000 | 4.036000 | 168.887000 | 237.400000 | 363.30000 |

8 rows × 27 columns

```
# Checking for correlation

plt.figure(figsize=(25, 12))
snsheatmap = sns.heatmap(df.corr(),annot = True, cmap ="Greens")
plt.show()
```

```
<ipython-input-33-1f1406066b4e>:4: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future versio
  snsheatmap = sns.heatmap(df.corr(),annot = True, cmap ="Greens")
```




Based on correlation matrix we prepared the pairplot of following column:

Crude Birth Rate: It's a fundamental demographic measure and seems to have varying degrees of correlation with other measures.

Total Fertility Rate: This is a key measure of fertility and shows the potential for future growth, which is not highly correlated with the crude birth rate.

Life Expectancy Female: Chosen over life expectancy male due to the social and health factors that might differ between genders.

Infant Mortality: A critical indicator of health and wellbeing in a country, and it's interesting to see how this relates to life expectancy and fertility rates.

Gross Reproduction Rate: It's related to fertility but focuses on female births, providing a different perspective on population replacement levels.

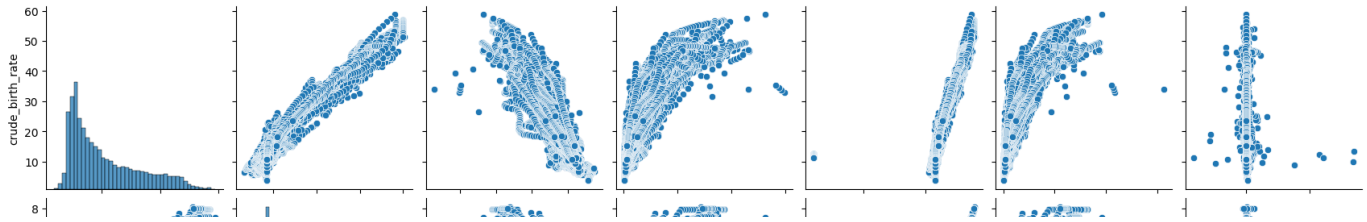
Mortality Rate Under 5: This provides additional insight into child survival and health beyond the first year (infant mortality).

Net Migration: Migration can significantly affect population structure and growth, and it has a low to moderate correlation with most of the other variables.

```
selected_columns = [
    'crude_birth_rate',
    'total_fertility_rate',
    'life_expectancy_female',
    'infant_mortality',
    'gross_reproduction_rate',
    'mortality_rate_under5',
    'net_migration'
]


sns.pairplot(df[selected_columns])
```

```
<seaborn.axisgrid.PairGrid at 0x7d545bba9990>
```




**Correlations:** Negative correlations are evident, such as between 'total_fertility_rate' and 'life_expectancy_female', indicating that as fertility decreases, female life expectancy tends to increase.

**Distributions:** Most variables are right-skewed, with a majority of countries having lower rates of demographic indicators like fertility and mortality.

**Outliers:** The variable 'net_migration' shows notable outliers, suggesting extreme migration scenarios for certain countries.

**Data Density:** Data points are densely clustered in moderate values for 'crude_birth_rate' and 'total_fertility_rate', reflecting commonality in reproductive rates across the dataset.

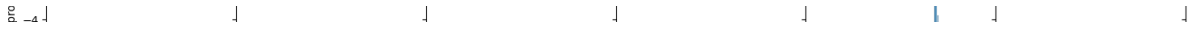
**Potential Issues:** Patterns of straight lines in plots involving 'net_migration' could point to data reporting issues or a large number of countries with no net migration.

**Variable Relationships:** The plot between 'infant_mortality' and 'life_expectancy_female' suggests a non-linear relationship, with life expectancy gains slowing as infant mortality approaches lower rates.

**Infant Mortality vs. Mortality Under 5:** The strong positive correlation between infant mortality and under-5 mortality reflects their shared influences on child health outcomes.

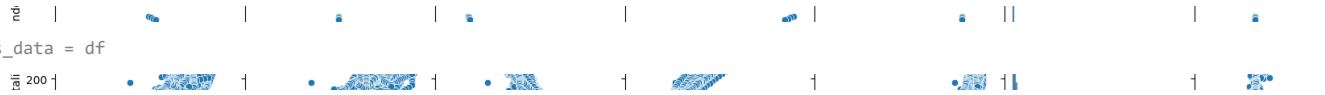
**Reproductive Indicators:** Fertility indicators correlate positively with one another, consistent with their common demographic influences.

**Migration:** The 'net_migration' variable displays weak and complex relationships with other demographic measures, implying diverse migration dynamics.



—

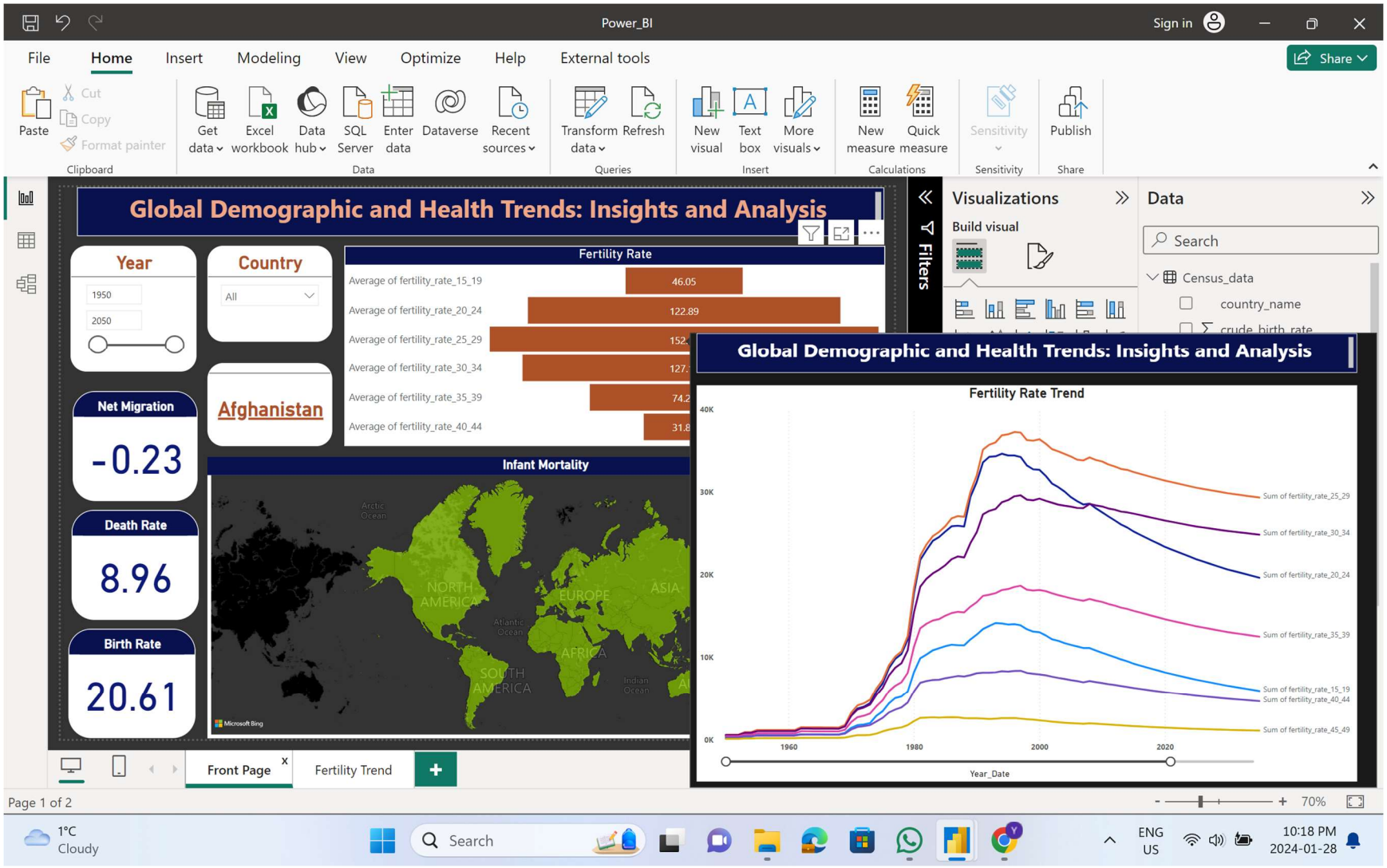
## Saving File to GDrive



```
Census_data = df
```



```
file_path = '/content/drive/MyDrive/Colab Notebooks/Project/Census_data.csv'
```


Power_BI
Sign in
File
Home
Insert
Modeling
View
Optimize
Help
External tools
Share
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Publish
Share
Global Demographic and Health Trends: Insights and Analysis
Year
1950
2050
Country
All
Afghanistan
Net Migration
-0.23
Death Rate
8.96
Birth Rate
20.61
Fertility Rate
Average of fertility_rate_15_19
46.05
Average of fertility_rate_20_24
122.89
Average of fertility_rate_25_29
Average of fertility_rate_30_34
Average of fertility_rate_35_39
74.2
Average of fertility_rate_40_44
31.8
Infant Mortality
Arctic Ocean
NORTH AMERICA
EUROPE
ASIA
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
Microsoft Bing
Filters
Visualizations
Build visual
Data
Search
Census_data
country_name
crude_birth_rate
Global Demographic and Health Trends: Insights and Analysis
Fertility Rate Trend
40K
30K
20K
10K
0K
1960
1980
2000
2020
Year_Date
Sum of fertility_rate_25_29
Sum of fertility_rate_30_34
Sum of fertility_rate_20_24
Sum of fertility_rate_35_39
Sum of fertility_rate_15_19
Sum of fertility_rate_40_44
Sum of fertility_rate_45_49
Front Page
Fertility Trend
Page 1 of 2
70%
1°C
Cloudy
Search
ENG
US
10:18 PM
2024-01-28