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()
```

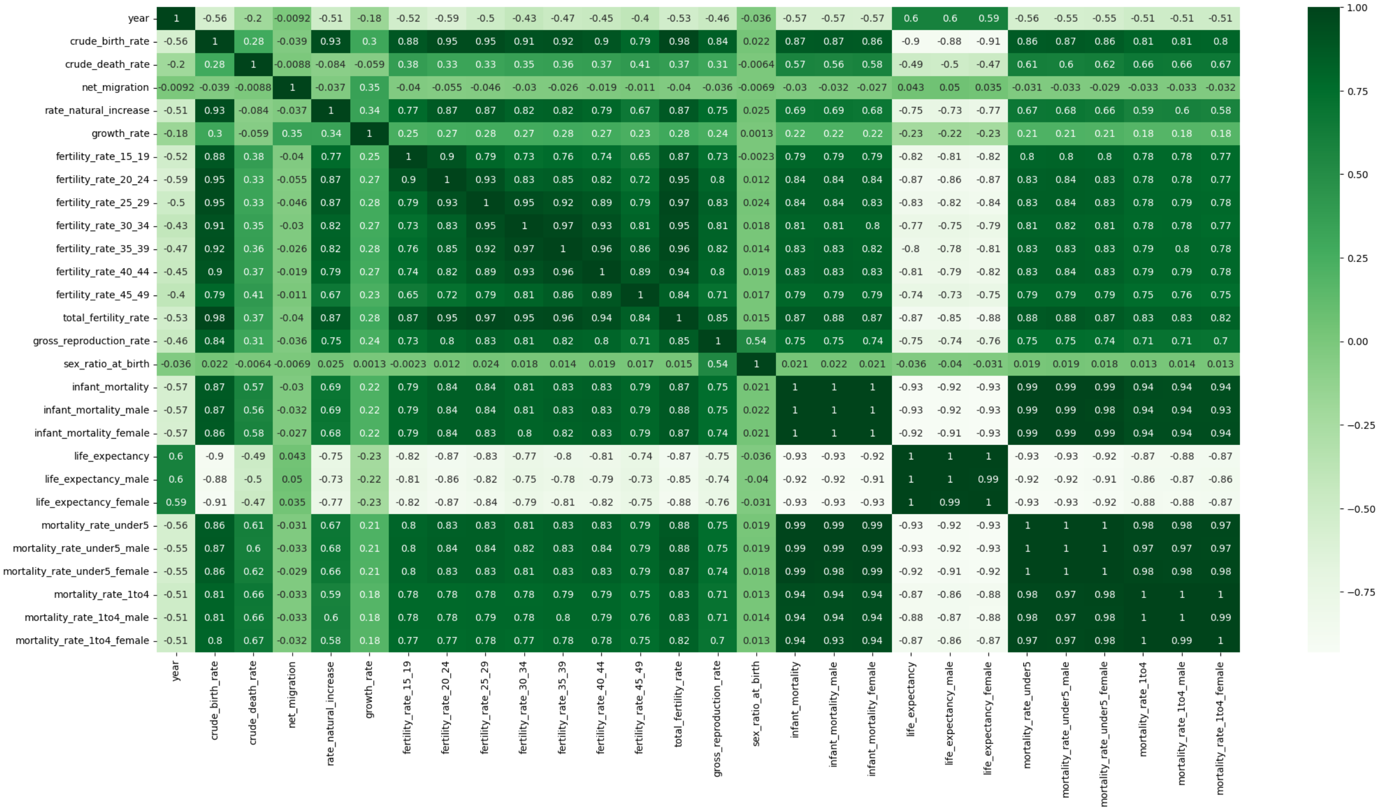
	crude_birth_rate	crude_death_rate	net_migration	rate_natural_increase	growth_rate	fertility_rate_15_19	fertility_rate_20_2
count	15016.000000	15016.000000	15016.000000	15016.000000	15016.000000	15016.000000	15016.00000
mean	20.605202	8.961052	-0.229858	1.164401	1.103004	46.048815	122.89374
std	11.705203	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")
```



	year	crude_birth_rate	crude_death_rate	net_migration	rate_natural_increase	growth_rate	fertility_rate_15_19	fertility_rate_20_24	fertility_rate_25_29	fertility_rate_30_34	fertility_rate_35_39	fertility_rate_40_44	fertility_rate_45_49	total_fertility_rate	gross_reproduction_rate	sex_ratio_at_birth	infant_mortality	infant_mortality_male	infant_mortality_female	life_expectancy	life_expectancy_male	life_expectancy_female	mortality_rate_under5	mortality_rate_under5_male	mortality_rate_under5_female	mortality_rate_1to4	mortality_rate_1to4_male	mortality_rate_1to4_female
year	1	-0.56	-0.2	-0.0092	-0.51	-0.18	-0.52	-0.59	-0.5	-0.43	-0.47	-0.45	-0.4	-0.53	-0.46	-0.036	-0.57	-0.57	-0.57	0.6	0.6	0.59	-0.56	-0.55	-0.55	-0.51	-0.51	-0.51
crude_birth_rate	-0.56	1	0.28	-0.039	0.93	0.3	0.88	0.95	0.95	0.91	0.92	0.9	0.79	0.98	0.84	0.022	0.87	0.87	0.86	-0.9	-0.88	-0.91	0.86	0.87	0.86	0.81	0.81	0.8
crude_death_rate	-0.2	0.28	1	-0.0088	-0.084	-0.059	0.38	0.33	0.33	0.35	0.36	0.37	0.41	0.37	0.31	-0.0064	0.57	0.56	0.58	-0.49	-0.5	-0.47	0.61	0.6	0.62	0.66	0.66	0.67
net_migration	-0.0092	-0.039	-0.0088	1	-0.037	0.35	-0.04	-0.055	-0.046	-0.03	-0.026	-0.019	-0.011	-0.04	-0.036	-0.0069	-0.03	-0.032	-0.027	0.043	0.05	0.035	-0.031	-0.033	-0.029	-0.033	-0.033	-0.032
rate_natural_increase	-0.51	0.93	-0.084	-0.037	1	0.34	0.77	0.87	0.87	0.82	0.82	0.79	0.67	0.87	0.75	0.025	0.69	0.69	0.68	-0.75	-0.73	-0.77	0.67	0.68	0.66	0.59	0.6	0.58
growth_rate	-0.18	0.3	-0.059	0.35	0.34	1	0.25	0.27	0.28	0.27	0.28	0.27	0.23	0.28	0.24	0.0013	0.22	0.22	0.22	-0.23	-0.22	-0.23	0.21	0.21	0.21	0.18	0.18	0.18
fertility_rate_15_19	-0.52	0.88	0.38	-0.04	0.77	0.25	1	0.9	0.79	0.73	0.76	0.74	0.65	0.87	0.73	-0.0023	0.79	0.79	0.79	-0.82	-0.81	-0.82	0.8	0.8	0.8	0.78	0.78	0.77
fertility_rate_20_24	-0.59	0.95	0.33	-0.055	0.87	0.27	0.9	1	0.93	0.83	0.85	0.82	0.72	0.95	0.8	0.012	0.84	0.84	0.84	-0.87	-0.86	-0.87	0.83	0.84	0.83	0.78	0.78	0.77
fertility_rate_25_29	-0.5	0.95	0.33	-0.046	0.87	0.28	0.79	0.93	1	0.95	0.92	0.89	0.79	0.97	0.83	0.024	0.84	0.84	0.83	-0.83	-0.82	-0.84	0.83	0.84	0.83	0.78	0.79	0.78
fertility_rate_30_34	-0.43	0.91	0.35	-0.03	0.82	0.27	0.73	0.83	0.95	1	0.97	0.93	0.81	0.95	0.81	0.018	0.81	0.81	0.8	-0.77	-0.75	-0.79	0.81	0.82	0.81	0.78	0.78	0.77
fertility_rate_35_39	-0.47	0.92	0.36	-0.026	0.82	0.28	0.76	0.85	0.92	0.97	1	0.96	0.86	0.96	0.82	0.014	0.83	0.83	0.82	-0.8	-0.78	-0.81	0.83	0.83	0.83	0.79	0.8	0.78
fertility_rate_40_44	-0.45	0.9	0.37	-0.019	0.79	0.27	0.74	0.82	0.89	0.93	0.96	1	0.89	0.94	0.8	0.019	0.83	0.83	0.83	-0.81	-0.79	-0.82	0.83	0.84	0.83	0.79	0.79	0.78
fertility_rate_45_49	-0.4	0.79	0.41	-0.011	0.67	0.23	0.65	0.72	0.79	0.81	0.86	0.89	1	0.84	0.71	0.017	0.79	0.79	0.79	-0.74	-0.73	-0.75	0.79	0.79	0.79	0.75	0.76	0.75
total_fertility_rate	-0.53	0.98	0.37	-0.04	0.87	0.28	0.87	0.95	0.97	0.95	0.96	0.94	0.84	1	0.85	0.015	0.87	0.88	0.87	-0.87	-0.85	-0.88	0.88	0.88	0.87	0.83	0.83	0.82
gross_reproduction_rate	-0.46	0.84	0.31	-0.036	0.75	0.24	0.73	0.8	0.83	0.81	0.82	0.8	0.71	0.85	1	0.54	0.75	0.75	0.74	-0.75	-0.74	-0.76	0.75	0.75	0.74	0.71	0.71	0.7
sex_ratio_at_birth	-0.036	0.022	-0.0064	-0.0069	0.025	0.0013	-0.0023	0.012	0.024	0.018	0.014	0.019	0.017	0.015	0.54	1	0.021	0.022	0.021	-0.036	-0.04	-0.031	0.019	0.019	0.018	0.013	0.014	0.013
infant_mortality	-0.57	0.87	0.57	-0.03	0.69	0.22	0.79	0.84	0.84	0.81	0.83	0.83	0.79	0.87	0.75	0.021	1	1	1	-0.93	-0.92	-0.93	0.99	0.99	0.99	0.94	0.94	0.94
infant_mortality_male	-0.57	0.87	0.56	-0.032	0.69	0.22	0.79	0.84	0.84	0.81	0.83	0.83	0.79	0.88	0.75	0.022	1	1	1	-0.93	-0.92	-0.93	0.99	0.99	0.98	0.94	0.94	0.93
infant_mortality_female	-0.57	0.86	0.58	-0.027	0.68	0.22	0.79	0.84	0.83	0.8	0.82	0.83	0.79	0.87	0.74	0.021	1	1	1	-0.92	-0.91	-0.93	0.99	0.99	0.99	0.94	0.94	0.94
life_expectancy	0.6	-0.9	-0.49	0.043	-0.75	-0.23	-0.82	-0.87	-0.83	-0.77	-0.8	-0.81	-0.74	-0.87	-0.75	-0.036	-0.93	-0.93	-0.92	1	1	1	-0.93	-0.93	-0.92	-0.87	-0.88	-0.87
life_expectancy_male	0.6	-0.88	-0.5	0.05	-0.73	-0.22	-0.81	-0.86	-0.82	-0.75	-0.78	-0.79	-0.73	-0.85	-0.74	-0.04	-0.92	-0.92	-0.91	1	1	0.99	-0.92	-0.92	-0.91	-0.86	-0.87	-0.86
life_expectancy_female	0.59	-0.91	-0.47	0.035	-0.77	-0.23	-0.82	-0.87	-0.84	-0.79	-0.81	-0.82	-0.75	-0.88	-0.76	-0.031	-0.93	-0.93	-0.93	1	0.99	1	-0.93	-0.93	-0.92	-0.88	-0.88	-0.87
mortality_rate_under5	-0.56	0.86	0.61	-0.031	0.67	0.21	0.8	0.83	0.83	0.81	0.83	0.83	0.79	0.88	0.75	0.019	0.99	0.99	0.99	-0.93	-0.92	-0.93	1	1	1	0.98	0.98	0.97
mortality_rate_under5_male	-0.55	0.87	0.6	-0.033	0.68	0.21	0.8	0.84	0.84	0.82	0.83	0.84	0.79	0.88	0.75	0.019	0.99	0.99	0.99	-0.93	-0.92	-0.93	1	1	1	0.97	0.97	0.97
mortality_rate_under5_female	-0.55	0.86	0.62	-0.029	0.66	0.21	0.8	0.83	0.83	0.81	0.83	0.83	0.79	0.87	0.74	0.018	0.99	0.98	0.99	-0.92	-0.91	-0.92	1	1	1	0.98	0.98	0.98
mortality_rate_1to4	-0.51	0.81	0.66	-0.033	0.59	0.18	0.78	0.78	0.78	0.78	0.79	0.79	0.75	0.83	0.71	0.013	0.94	0.94	0.94	-0.87	-0.86	-0.88	0.98	0.97	0.98	1	1	1
mortality_rate_1to4_male	-0.51	0.81	0.66	-0.033	0.6	0.18	0.78	0.78	0.79	0.78	0.8	0.79	0.76	0.83	0.71	0.014	0.94	0.94	0.94	-0.88	-0.87	-0.88	0.98	0.97	0.98	1	1	0.99
mortality_rate_1to4_female	-0.51	0.8	0.67	-0.032	0.58	0.18	0.77	0.77	0.78	0.77	0.78	0.78	0.75	0.82	0.7	0.013	0.94	0.93	0.94	-0.87	-0.86	-0.87	0.97	0.97	0.98	1	0.99	1

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.

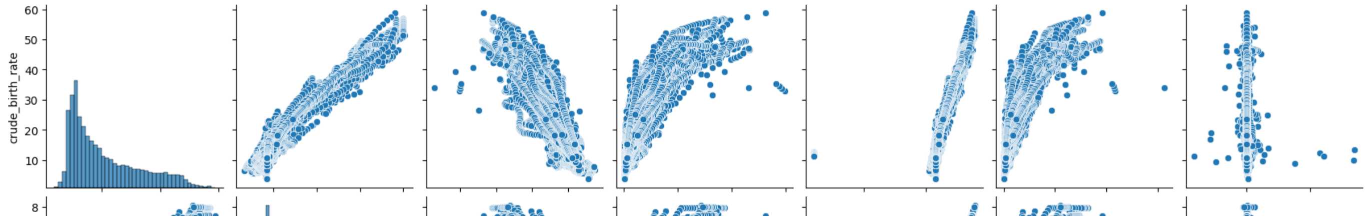
Mortality Rate Under 5: This provides additional insight into child survival and health beyond the first year (infant mortality).

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.

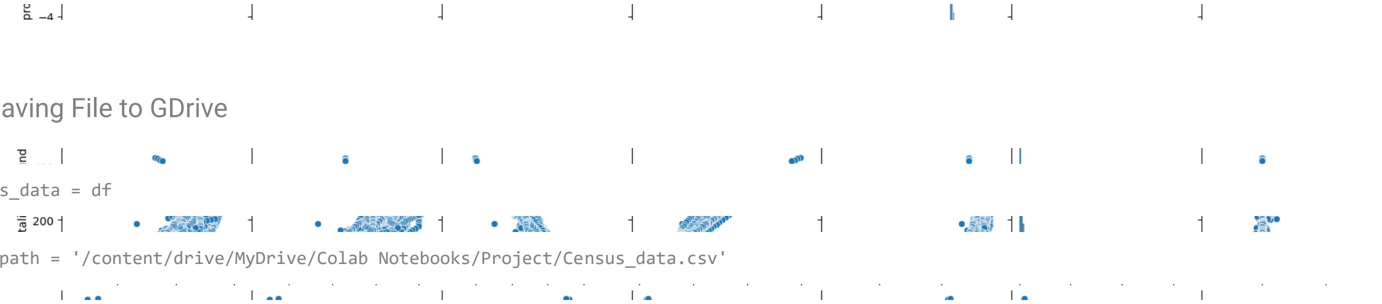
Potential Issues: Patterns of straight lines in plots involving 'net_migration' could point to data reporting issues or a large number of countries with no net migration.

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'
```

```
Census_data.to_csv(file_path, index=False)
```