Data Visualization

Exploratory Data Analysis

—

INTRODUCTION

The Human Development Index (HDI) is a summary measure of achievements in key dimensions of human development: a long and healthy life, access to knowledge and a decent standard of living. There are six main factors for Human Development that play an important role to distinguish among first world countries, countries under development and third world countries. These factors include gender development, gender inequality, historical index, human development, inequality adjusted, and multidimensional poverty.

For our Analysis We chose the report *'Gender Inequality'* and *'Gender Development'*. First Report emphasizes the effect of gender inequality on HDI (Human Development Index) and GII (Gender Development Index), while the other report measures gender inequalities in achievement in three basic dimensions of human development (health, education, and Economic Resource Contribution).

'Understanding the dataset' can refer to a number of things including but not limited to Extracting important variables and leaving behind useless variables Identifying outliers, missing values, or human error. Understanding the relationship(s), or lack of, between variables

Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process.

- Mount Drive and Importing Libraries

We mounted the google drive in order to load the reports *'gender_inequality.csv'* and *'gender_development.csv'* . Also we imported libraries such as *pandas, numpy, matplotlib* and *plotly* as a project requirements.

```
[2]  1  from google.colab import drive
     2  drive.mount(r"/content/drive")
     3  %cd '/content/drive/MyDrive/Colab Notebooks/Data Visualization Assignments/Final Project/'

Mounted at /content/drive
/content/drive/MyDrive/Colab Notebooks/Data Visualization Assignments/Final Project

[3]  1  import pandas as pd
     2  import numpy as np
     3  import matplotlib.pyplot as plt
     4  import seaborn as sns
     5  sns.set(rc={'figure.figsize':(12,8)})
     6  pd.set_option('display.precision',2)
     7  import plotly.offline as py
     8  import plotly.express as px
     9  import plotly.graph_objs as go
    10  #import cufflinks as cf
    11  from mpl_toolkits.mplot3d import Axes3D
    12  import matplotlib as mpl
    13
```

- Loading and Summary of the Dataset

Finally importing the csv files for our analysis. We can notice both reports have the same number of records where each record is based on a particular country. Therefore It is possible to merge both reports based on Country names.

```
[ ]  1  #reading the data
     2
     3  #uploaded = files.upload()
     4  df1 = pd.read_csv("gender_inequality.csv")
     5  df2=pd.read_csv("gender_development.csv")

[ ]  1  print(df1.shape, df2.shape)

(195, 10) (195, 13)
```

Summary of the dataset with respect to Gender Inequality

```
1  df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 10 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   GII Rank                                188 non-null    float64
 1   Country                                 195 non-null    object
 2   Gender Inequality Index (GII)           195 non-null    object
 3   Maternal Mortality Ratio                195 non-null    object
 4   Adolescent Birth Rate                   195 non-null    object
 5   Percent Representation in Parliament    195 non-null    object
 6   Population with Secondary Education (Female)  195 non-null  object
 7   Population with Secondary Education (Male)    195 non-null  object
 8   Labour Force Participation Rate (Female)      195 non-null  object
 9   Labour Force Participation Rate (Male)        195 non-null  object
dtypes: float64(1), object(9)
memory usage: 15.4+ KB
```

Gender Development Index (GDI)	Human Development Index (Female)	Human Development Index (Male)	Life Expectancy at Birth (Female)	Life Expectancy at Birth (Male)	Expected Years of Education (Female)	Expected Years of Education (Male)	Mean Years of Education (Female)	Mean Years of Education (Male)	Estimated Gross National Income per Capita (Female)	Estimated Gross National Income per Capita (Male)
0.996	0.94	0.944	83.6	79.5	18.2	16.8	12.7	12.5	57140	72825
0.976	0.922	0.945	84.5	80.3	20.7	19.7	13.1	12.9	33688	50914
0.95	0.898	0.945	85	80.8	15.7	15.9	11.5	13.1	44132	69077
0.977	0.912	0.934	82.2	78.3	19.3	18.1	12.8	12.7	36439	51727
0.947	0.893	0.943	83.3	79.7	18	17.9	11.6	12.2	29500	61641

Summary of the dataset with respect to Gender Development

```
1  df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 13 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   GDI Rank                                        188 non-null    float64
 1   Country                                         195 non-null    object
 2   Gender Development Index (GDI)                  195 non-null    object
 3   Human Development Index (Female)                195 non-null    object
 4   Human Development Index (Male)                  195 non-null    object
 5   Life Expectancy at Birth (Female)               195 non-null    object
 6   Life Expectancy at Birth (Male)                 195 non-null    object
 7   Expected Years of Education (Female)            195 non-null    object
 8   Expected Years of Education (Male)              195 non-null    object
 9   Mean Years of Education (Female)                195 non-null    object
 10  Mean Years of Education (Male)                  195 non-null    object
 11  Estimated Gross National Income per Capita (Female)  195 non-null  object
 12  Estimated Gross National Income per Capita (Male)    195 non-null  object
dtypes: float64(1), object(12)
memory usage: 19.9+ KB
```

Country	Gender Development Index (GDI)	Human Development Index (Female)	Human Development Index (Male)	Life Expectancy at Birth (Female)	Life Expectancy at Birth (Male)	Expected Years of Education (Female)	Expected Years of Education (Male)	Mean Years of Education (Female)	Mean Years of Education (Male)	Estimated Gross National Income per Capita (Female)	Estimated Gross National Income per Capita (Male)
Norway	0.996	0.94	0.944	83.6	79.5	18.2	16.8	12.7	12.5	57140	72825
Australia	0.976	0.922	0.945	84.5	80.3	20.7	19.7	13.1	12.9	33688	50914
witzerland	0.95	0.898	0.945	85	80.8	15.7	15.9	11.5	13.1	44132	69077
Denmark	0.977	0.912	0.934	82.2	78.3	19.3	18.1	12.8	12.7	36439	51727
etherlands	0.947	0.893	0.943	83.3	79.7	18	17.9	11.6	12.2	29500	61641

- **Merging both the Dataset**

Since both of our tables have the same numbers of rows or records, we can merge both the tables based on axis=1 or features.

```
[ ]  1
     2  df=df1.merge(df2)
     3  #The variable csv_file is going to be a string object that points to the location of our data
     4  #the next line asks our pandas library to try to read our CSV file, and turn it into a dataframe.
     5  #Once executed, it will be stored in the variable df.
```

```
[ ]  1  df.shape
```

```
(195, 22)
```

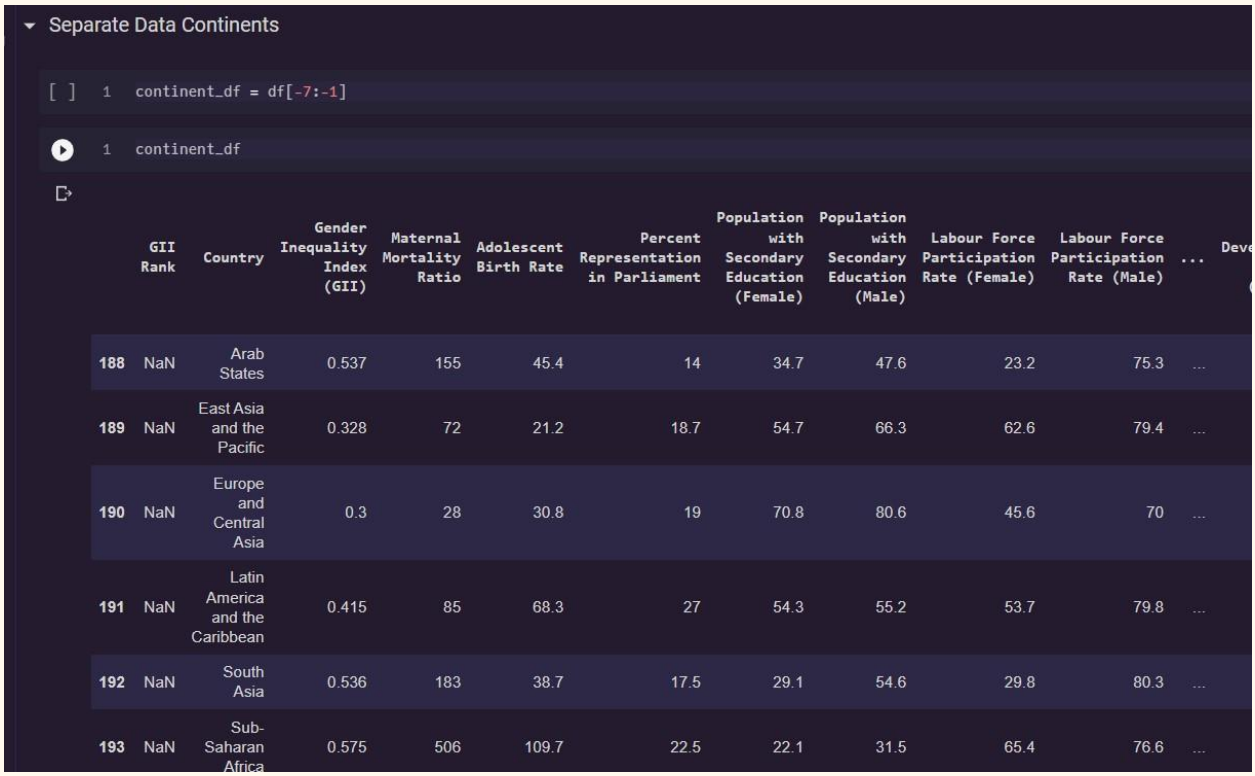
```
[ ]  1  df.head()
```

	GII Rank	Country	Gender Inequality Index (GII)	Maternal Mortality Ratio	Adolescent Birth Rate	Percent Representation in Parliament	Population with Secondary Education (Female)	Population with Secondary Education (Male)	Labour Force Participation Rate (Female)	Labour Force Participation Rate (Male)	...	Develc (Fe
0	1.0	Norway	0.067	4	7.8	39.6	97.4	96.7	61.2	68.7	...	
1	2.0	Australia	0.11	6	12.1	30.5	94.3	94.6	58.8	71.8	...	
2	3.0	Switzerland	0.028	6	1.9	28.5	95	96.6	61.8	74.9	...	
3	4.0	Denmark	0.048	5	5.1	38	95.5	96.6	58.7	66.4	...	
4	5.0	Netherlands	0.062	6	6.2	36.9	87.7	90.5	58.5	70.6	...	

5 rows × 22 columns

- **Separating Features for continent based data**

We can observe the last 5 records in our merged data is not for countries but based on Continents (Arabs, Sub Saharan African, East Asia, South Asia, Europe, and South America). Therefore, we'll drop them and build a new dataframe off of it. This will help us understand the summary in terms of continents.



Separate Data Continents

```
[ ]  1  continent_df = df[-7:-1]
```

```
1  continent_df
```

	GII Rank	Country	Gender Inequality Index (GII)	Maternal Mortality Ratio	Adolescent Birth Rate	Percent Representation in Parliament	Population with Secondary Education (Female)	Population with Secondary Education (Male)	Labour Force Participation Rate (Female)	Labour Force Participation Rate (Male)	...	Deve
188	NaN	Arab States	0.537	155	45.4	14	34.7	47.6	23.2	75.3	...	
189	NaN	East Asia and the Pacific	0.328	72	21.2	18.7	54.7	66.3	62.6	79.4	...	
190	NaN	Europe and Central Asia	0.3	28	30.8	19	70.8	80.6	45.6	70	...	
191	NaN	Latin America and the Caribbean	0.415	85	68.3	27	54.3	55.2	53.7	79.8	...	
192	NaN	South Asia	0.536	183	38.7	17.5	29.1	54.6	29.8	80.3	...	
193	NaN	Sub-Saharan Africa	0.575	506	109.7	22.5	22.1	31.5	65.4	76.6	...	

- **Data Preparation**

This step is performed after the data gathering procedure. Our dataset is organized already so we do not need to work on building a data pipeline to restructure the data. However, our data set has certain anomalies that need to be taken care of.

1. Data Cleaning

This is the initial and most common task in data preparation that is performed on raw data. Data cleansing is the process of examining, identifying, and mitigating errors in raw data. Normally, the raw data are neither sufficiently complete nor sufficiently clean to directly train the ML model. Manually entered data can have incomplete, duplicated, erroneous, or inaccurate values.

*In our note book we removed "**incompleteness error"** by replace null values from the features **GDI Rank** and **GII Rank** with mean values.*

```
[ ]  1  # Checking null values
     2  df.isnull().sum()
```

```
GII Rank                                          7
Country                                           0
Gender Inequality Index (GII)                     0
Maternal Mortality Ratio                          0
Adolescent Birth Rate                             0
Percent Representation in Parliament              0
Population with Secondary Education (Female)      0
Population with Secondary Education (Male)        0
Labour Force Participation Rate (Female)          0
Labour Force Participation Rate (Male)            0
GDI Rank                                          7
Gender Development Index (GDI)                    0
Human Development Index (Female)                  0
Human Development Index (Male)                    0
Life Expectancy at Birth (Female)                 0
Life Expectancy at Birth (Male)                   0
Expected Years of Education (Female)              0
Expected Years of Education (Male)                0
Mean Years of Education (Female)                  0
Mean Years of Education (Male)                    0
Estimated Gross National Income per Capita (Female)    0
Estimated Gross National Income per Capita (Male)      0
dtype: int64
```

```
1  for i in df.columns:
2    if df[i].dtypes=='float':
3      df[i]=df[i].fillna(df[i].mean())
4
```

+ Code + Text

```
1  df.isnull().sum()
```

```
GII Rank                                                0
Country                                                 0
Gender Inequality Index (GII)                           0
Maternal Mortality Ratio                                0
Adolescent Birth Rate                                   0
Percent Representation in Parliament                    0
Population with Secondary Education (Female)            0
Population with Secondary Education (Male)              0
Labour Force Participation Rate (Female)                0
Labour Force Participation Rate (Male)                  0
GDI Rank                                                0
Gender Development Index (GDI)                          0
Human Development Index (Female)                        0
Human Development Index (Male)                          0
Life Expectancy at Birth (Female)                       0
Life Expectancy at Birth (Male)                         0
Expected Years of Education (Female)                    0
Expected Years of Education (Male)                      0
Mean Years of Education (Female)                        0
Mean Years of Education (Male)                          0
Estimated Gross National Income per Capita (Female)     0
Estimated Gross National Income per Capita (Male)       0
dtype: int64
```

2. Format Data

In this step, we format the data in order to facilitate visual encoding and modelling tasks. In our case we have formatted the data type of features of *'object'* type to *'float'* type.

Changing Data Type to Float

```
[ ]  1  df[df=='..']=np.nan
```

```
[ ]  1  country=df['Country']
     2  df_=df.drop(['Country'], axis=1)
     3  for i in df_.columns:
     4    if df_[i].dtypes!='float':
     5      df_[i]=df_[i].astype('float')
     6
```

```
[ ]  1  df_['Country']=country
```

However changing the data type also helped us find records which are invalid and null.

```
1  df_.isnull().sum()
```

```
GII Rank                                              0
Gender Inequality Index (GII)                        33
Maternal Mortality Ratio                             10
Adolescent Birth Rate                                 5
Percent Representation in Parliament                  3
Population with Secondary Education (Female)         26
Population with Secondary Education (Male)           26
Labour Force Participation Rate (Female)             11
Labour Force Participation Rate (Male)               11
GDI Rank                                              0
Gender Development Index (GDI)                       27
Human Development Index (Female)                     27
Human Development Index (Male)                       27
Life Expectancy at Birth (Female)                     5
Life Expectancy at Birth (Male)                       5
Expected Years of Education (Female)                 14
Expected Years of Education (Male)                   14
Mean Years of Education (Female)                     18
Mean Years of Education (Male)                       18
Estimated Gross National Income per Capita (Female)  11
Estimated Gross National Income per Capita (Male)    11
Country                                               0
dtype: int64
```

Therefore, we'll impute the missing data with mean values.

As we can see many values are not covertable and resulted in Nan (not a Number), therefore replacing them either with Mean or removing the records

```
for i in df_.columns:
  if df_[i].dtypes=='float':
    df_[i]=df_[i].fillna(df_[i].mean())
```

```
df_.isnull().sum()
```

```
GII Rank                                          0
Gender Inequality Index (GII)                     0
Maternal Mortality Ratio                          0
Adolescent Birth Rate                             0
Percent Representation in Parliament              0
Population with Secondary Education (Female)      0
Population with Secondary Education (Male)        0
Labour Force Participation Rate (Female)          0
Labour Force Participation Rate (Male)            0
GDI Rank                                          0
Gender Development Index (GDI)                    0
Human Development Index (Female)                  0
Human Development Index (Male)                    0
Life Expectancy at Birth (Female)                 0
Life Expectancy at Birth (Male)                   0
Expected Years of Education (Female)              0
Expected Years of Education (Male)                0
Mean Years of Education (Female)                  0
Mean Years of Education (Male)                    0
Estimated Gross National Income per Capita (Female)  0
Estimated Gross National Income per Capita (Male)    0
Country                                           0
dtype: int64
```

3. Scaling and Normalizing the Dataset

Scaling and Normalizing the Dataset

Since the distribution of our dataset for some vairable or features is skewed we are performing min-max scaling (similar to normalization) and shift methods to see its impacts

```
sub_df = df_.select_dtypes(include='float')

from sklearn.preprocessing import MinMaxScaler

ss = MinMaxScaler()
dfnorm = ss.fit_transform(sub_df)
dfnormfloat = pd.DataFrame(dfnorm, columns=sub_df.columns)
dfnorm = pd.concat([dfnormfloat, df_['Country']], axis=1)
dfnorm.head()
```

In this step the data has been normalize by minmaxscaler method.

- **Visual Encodings**

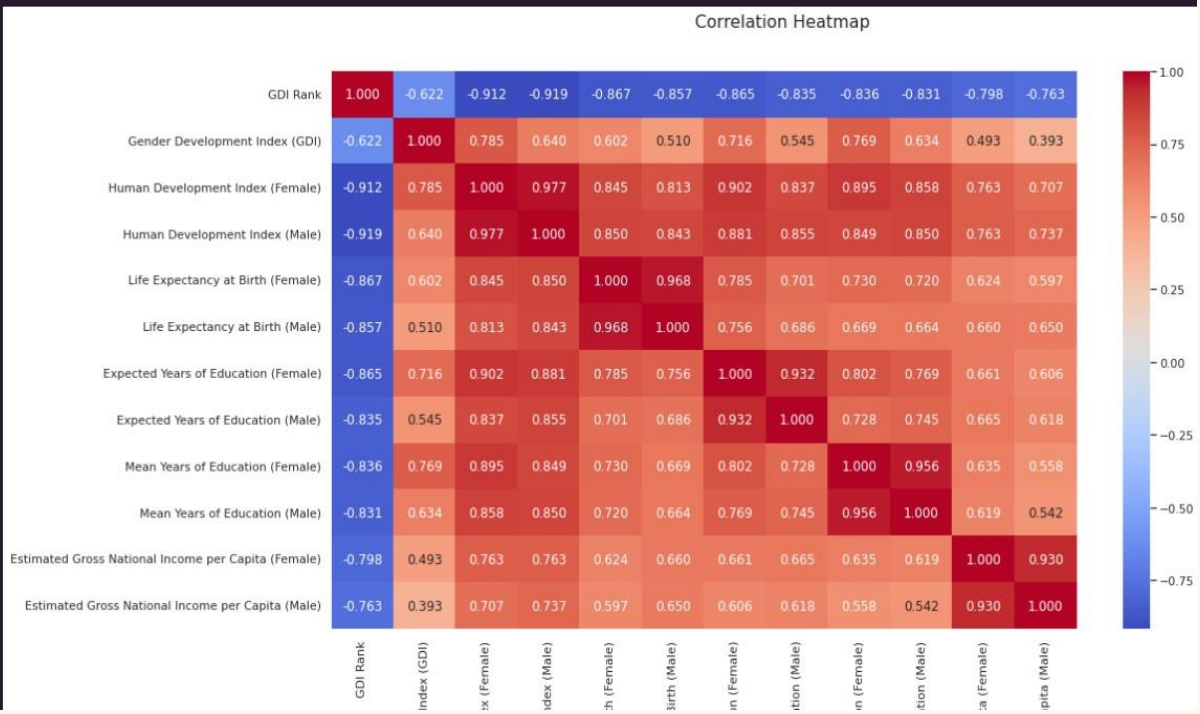
<u>Plot Selection 1 --Correlation Heatmap Matrix</u>

We have plot a heatmap to understand the correlation among various features, we can see some of the features are highly correlated, and generally they are the same features distributed among Male and Female.

As it's illustrated, all the variables are related to each other and since the correlation of all of them is above 0.5, then they have a high and moderate correlation, some of them have a positive relationship and some have a negative relationship.

Gender Development Correlation Matrix After Scaling

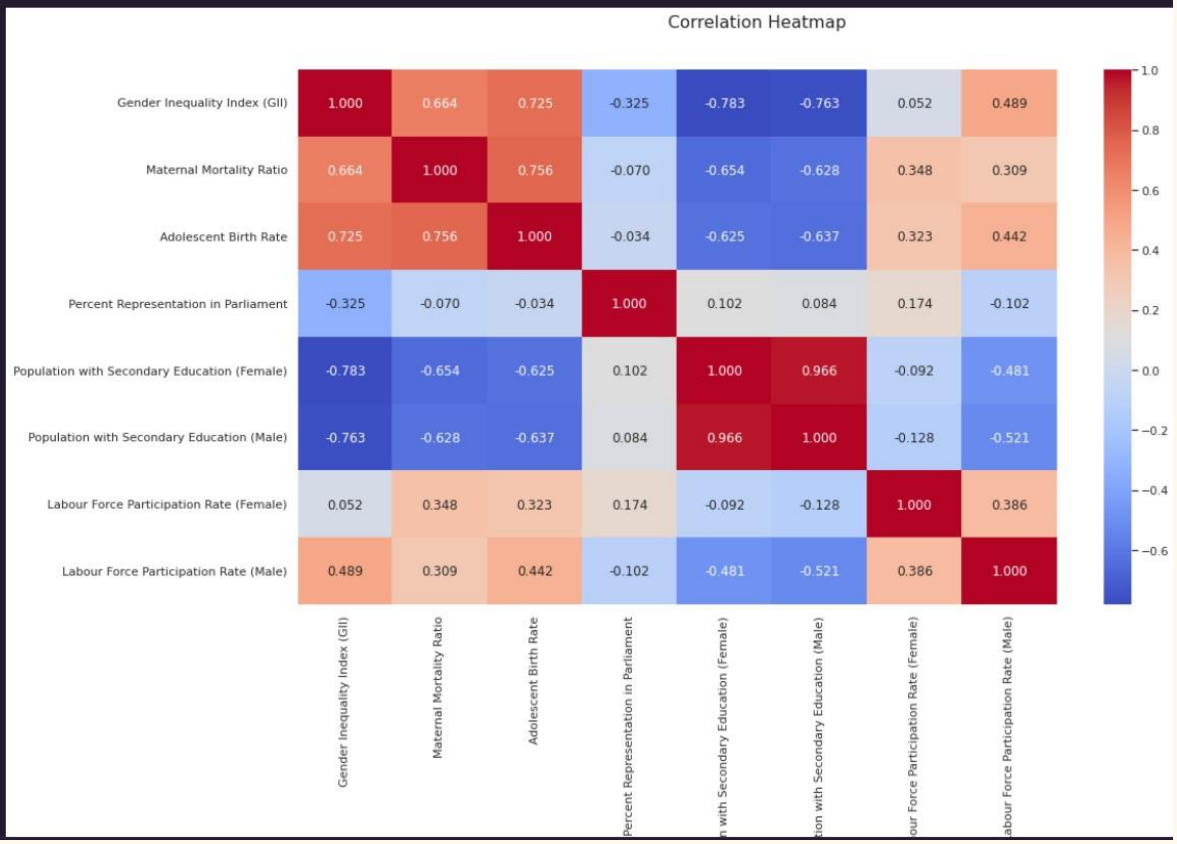
```
#corrolation heatmap for gender_development variable
fig, ax = plt.subplots(figsize=(16, 9))
corr = dfnormfloat.iloc[:,9:].corr()
htmap = sns.heatmap(corr, annot=True, ax=ax, cmap="coolwarm",fmt='.3f')
fig.subplots_adjust(top=0.90)
c= fig.suptitle('Correlation Heatmap', fontsize=16)
```



Correlation between Attributes of Gender development

Gender Inequality Correlation Matrix after Scaling

```
#corrolation heatmap  for Gender Inequality
fig, ax = plt.subplots(figsize=(16, 9))
corr = dfnormfloat.iloc[:,1:9].corr()
htmap = sns.heatmap(corr, annot=True, ax=ax, cmap="coolwarm",fmt='.3f')
fig.subplots_adjust(top=0.90)
c= fig.suptitle('Correlation Heatmap', fontsize=16)
```



Correlation between attributes of Gender Inequality

Plot Selection 2 - Interaction 3D Scatter Plot (with colour and size)

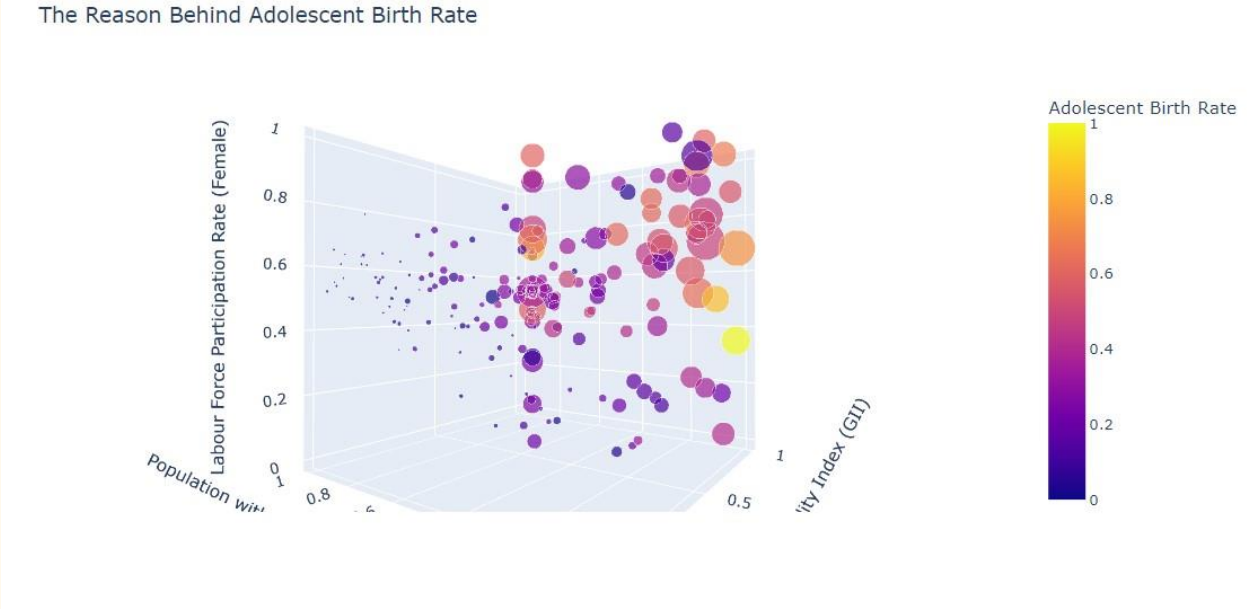
For the second visualization based on gender inequality, we decided to make an interactive 3D scatter plot (with colour and size as extra dimensions). This plot shows the relation between female labour force rate, female population with secondary education, gender inequality index (GII), adolescent birth rate as a colour and maternal mortality ratio as a size. The interactivity and the 5D nature of the plot blend well to analysing these variables, as the user can freely

move the plot around and hover over any point to see its exact values. The colour gradient from purple to yellow makes the data visually distinct and easy to parse.

- From the plot, we can perceive that with increasing female secondary education population, the gender inequality index trends towards 1. As a result, we can presume that a higher educated female population contributes to more gender equality.
- We can also observe that as the female secondary education population increase, the female labour force rate and GII increases, while the maternal mortality ratio and adolescent birth rate decreases. This is to be expected, as educating the female population and creating an environment of gender equality increases awareness of maternal complications/contraception and better equips the female population, causing less adolescent pregnancies and less general maternal mortalities.
- We can also observe a vertical line of outlier data in one of the vertices of the plot which muddies the data. This plot allows us to visualize all our variables in relation to the female population and determine that education is a great contributor to a lowered GII, despite the outlier data.

```python
#preparing data for plotly

data1 = px.scatter_3d(
    dfnorm,
    x='Gender Inequality Index (GII)',
    y='Population with Secondary Education (Female)',
    z='Labour Force Participation Rate (Female)',
    color='Adolescent Birth Rate',
    size='Maternal Mortality Ratio',
    title='The Reason Behind Adolescent Birth Rate',
    size_max=50,
    )

data = data1
layout = go.Layout(
    title='The Reason Behind Adolescent Birth Rate',
    scene=dict(
        xaxis=dict(
            title='Gender Inequality Index (GII)'
        ),
        yaxis=dict(
            title='Population With Secondary Education (Female)'
        ),
        zaxis=dict(
            title='Labour Force Participation Rate (Female)'
        )
    ),
    margin=dict(
        l=0,
        r=0,
        b=0,
        t=0
    )
)
fig = go.Figure(data=data, layout=layout)
fig.show()
```

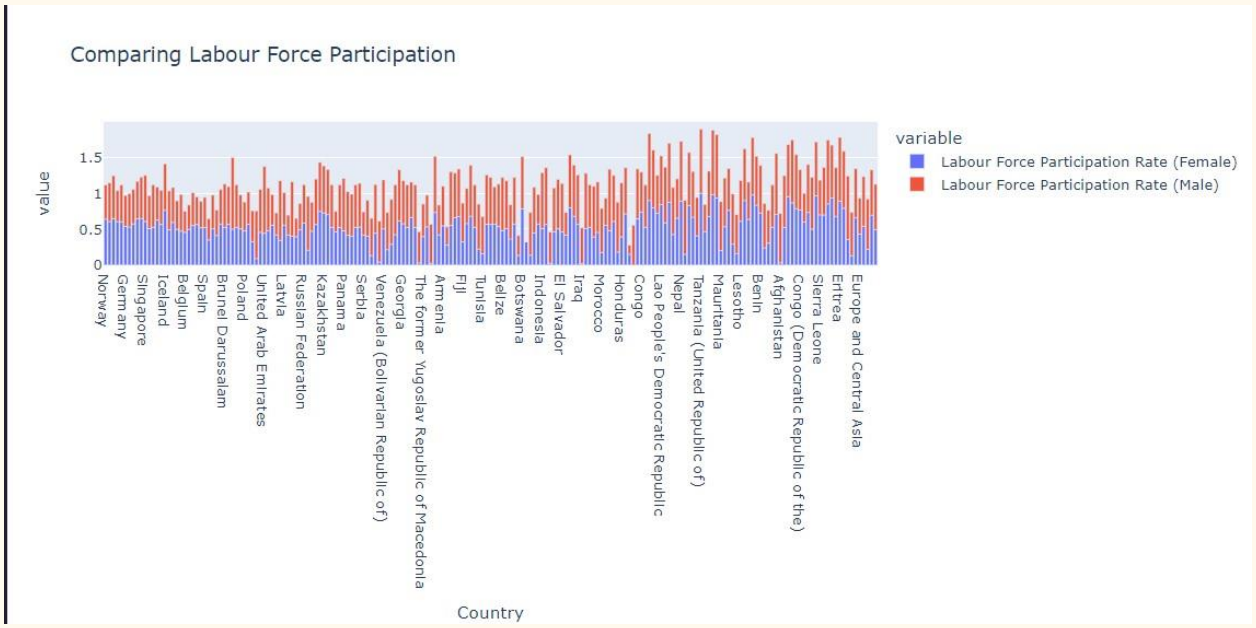
--NORMAL--



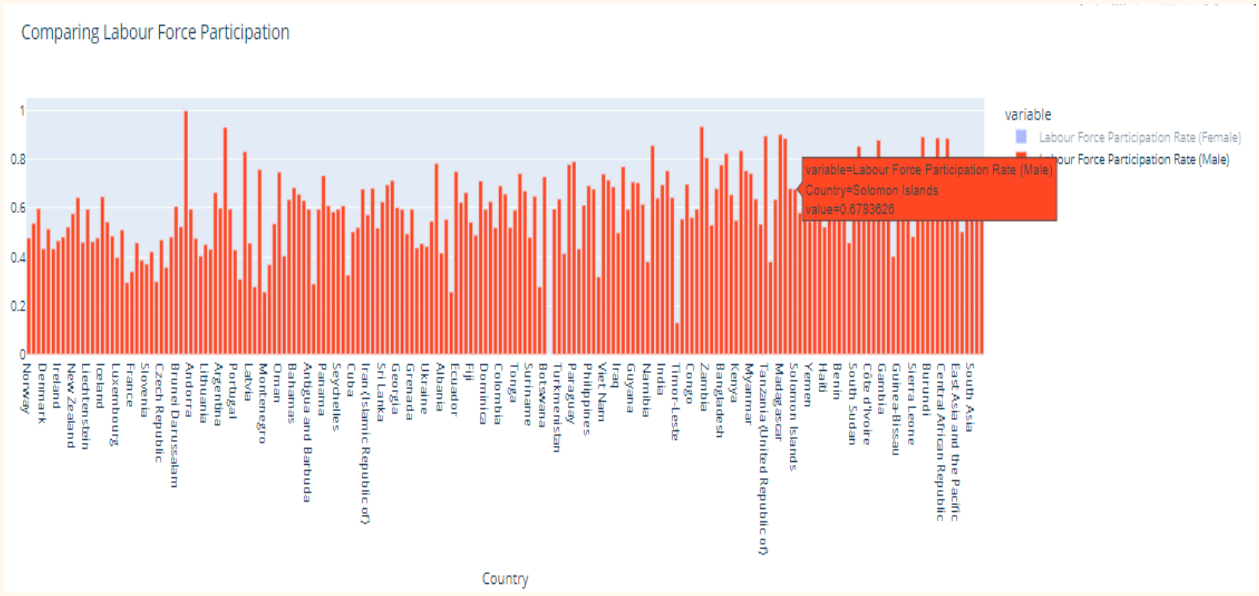
Plot Selection 3 -Interaction Bar Plot for Labour Participation

- The Bar Plot exhibit the proportion of Labour Force Participation in various countries by Male and Female
- It is evident that Men have a major contribution, while in some countries females have no contribution at all. which eventually leads to rise in Gender Inequality. Countries with almost equal Labour Force contribution are also doing good in terms of *'Gender Inequality Index'*.

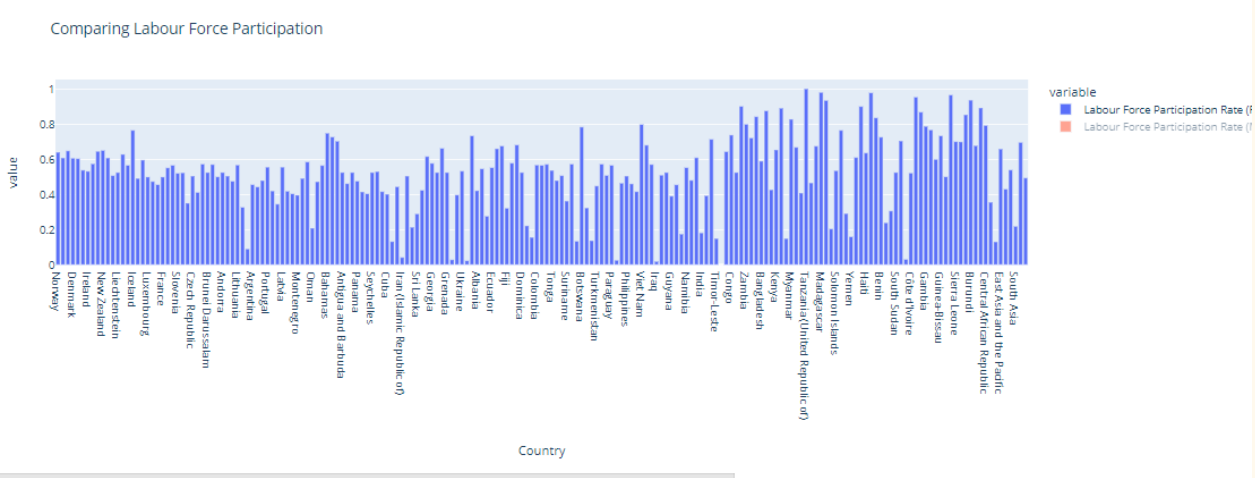
```
fig = px.bar(dfnorm,x='Country', y=['Labour Force Participation Rate (Male)','Labour Force Participation Rate (Female)'] , labe

fig.show()
```



- We can filter our plot by selecting the desired legend so that the plot is displayed only based on male or female.



In this case, we can clearly see that in the country of Moldova only women and in the country of Syrian Arab Republic only men participated in the labour force. According the above plot the most labour force participation among women belongs to Tanzania.

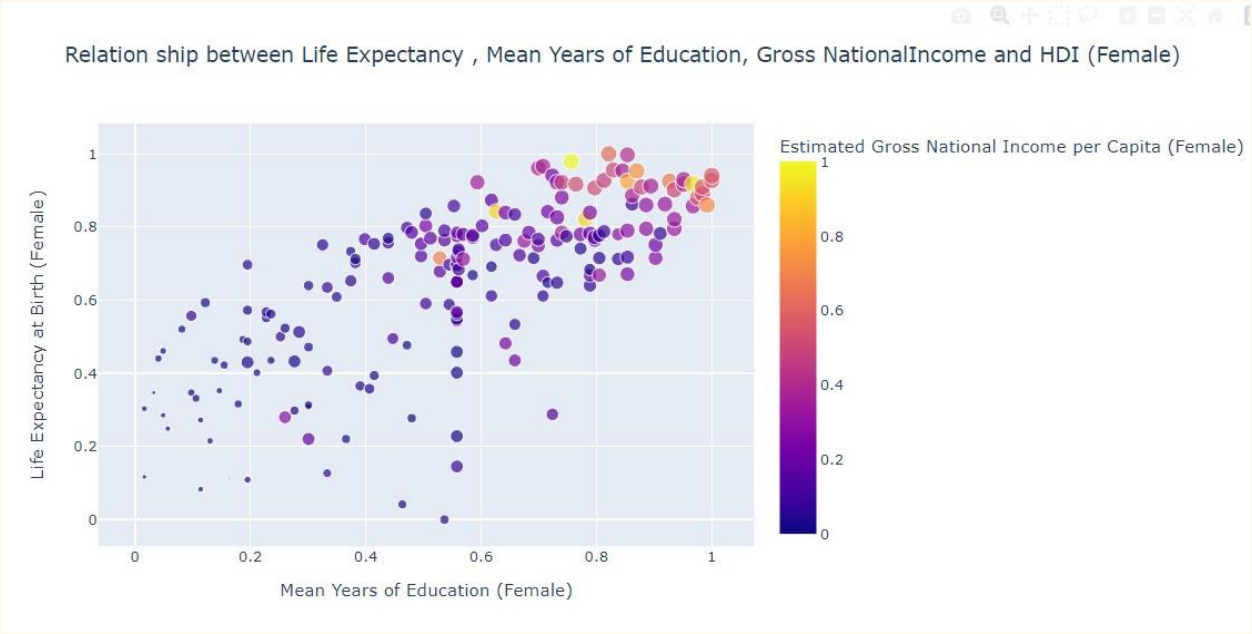


According the below plot the most labour force participation among men belongs to Qatar.

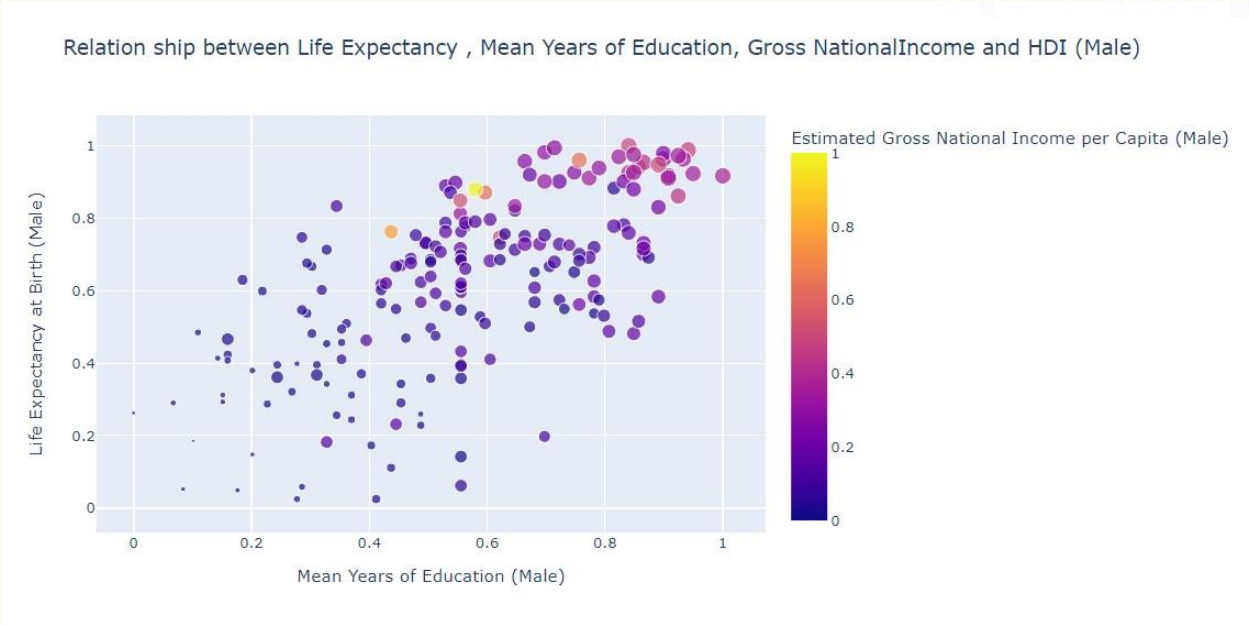
Plot 4 -Interaction 2D Scatter Plot(with color and size)

- We used Scatter plot to evaluate the relation of 'Mean years of Education' and 'Life Expectancy at Birth' on 'Estimate GNI' and 'HDI' for both male and female.
- Years of Education has direct a relationship with Life Expectancy
- The size of dot represents the Human Development Index, and it increases with the growth of GNI (as a colour) and Mean years of education, irrespective of the gender. So Human Development Index have a direct relationship with GNI and Mean years of education.
- It can be said that Life Expectancy and the Years of Education have a direct relationship with the Gross National Income as well as the Human Development Index.

```
dfnorm.sort_values(by=['Mean Years of Education (Female)', 'Life Expectancy at Birth (Female)',
                       'Estimated Gross National Income per Capita (Female)', 'Human Development Index (Female)'], inplace=True)


fig=px.scatter(dfnorm, x="Mean Years of Education (Female)", y="Life Expectancy at Birth (Female)", color="Estimated Gross Nati
                       size="Human Development Index (Female)",title='Relation ship between Life Expectancy , Mean Years of Educatio


fig.show()
```



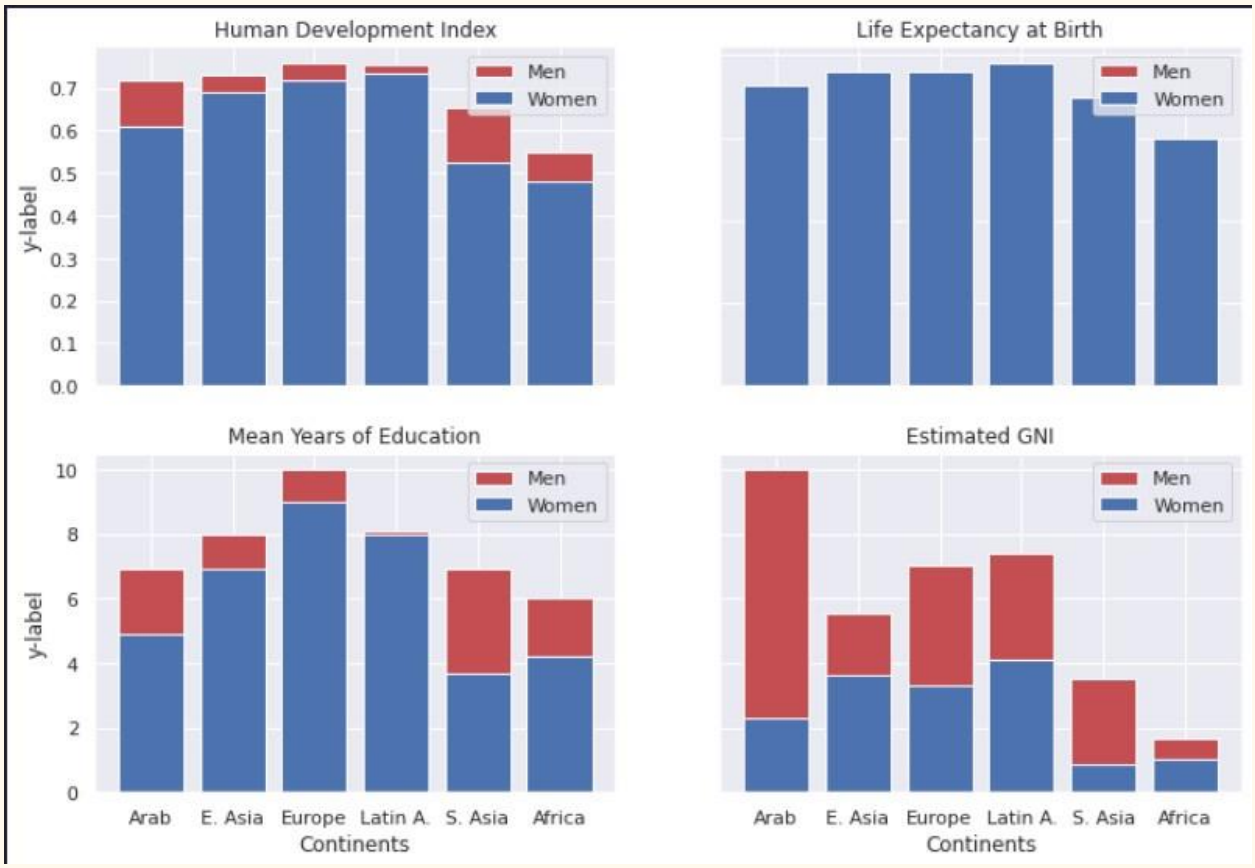

```
fig=px.scatter(dfnorm, x="Mean Years of Education (Male)",
               y="Life Expectancy at Birth (Male)",
               color="Estimated Gross National Income per Capita (Male)", size="Human Development Index (Male)",title='Relati
fig.show()

```



Plot Selection 5 - Bar Plot Comparison in Various Continents

- The Bar Plot exhibits the comparison between male and female in 5 different continents in terms of Human Development, Education, in Life expectancy, and Gross National Income.
- Women have significantly done better in Life expectancy at Birth in comparison to Men.
- On the other Hand, Male by a decent margin have better Gross National Income per capita than Female, which one of the contributing factor for inequality and gender biases
- In terms of “Mean years of Education”, It can be said that in South Asia, women spend about half the time of men in education. This is despite the fact in other continent women by a decent margin have high level of Mean years of Education than men.
- In the “Human Development Index”, women by a decent margin have better performance than men.

-


Human Development Index
Life Expectancy at Birth
Mean Years of Education
Estimated GNI
Men
Women
y-label
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
Arab
E. Asia
Europe
Latin A.
S. Asia
Africa
Continents

—---------- END ----------