Assignment 3 - Multivariate Statistics

Multivariate analysis of gender-wise labour force participation rates

Amrutha Seshagiri
Ayush Mishra

Indian Institute of Management Bangalore

1 *Introduction*

Our dataset is the country wise female and male labour force participation rates. This dataset includes the following variables - country code, country, continent, hemisphere, Human Development Index(HDI) Rank in 2021, labour force participation rates from 1990 to 2021. The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

In addition to this, we have included some more data at the country level, such as the HDI index, Gross Domestic Product (GDP), Fertility Rate (births per woman) and the percentage contribution to GDP by the following sectors - Industry, Agriculture, Services. We have limited our study to only these three sectors as these are the major sectors accounting for close to 95% of the GDP of most countries. All these datapoints were taken for the year 2021 to correspond to the year of the HDI rank in the original dataset.

Our rationale behind including these variables were:

- HDI Index - to have a continuous variable for meaningful modeling and interpretation
- GDP - to study the economic implications of labour force participation rates
- Fertility Rate - to check for relationship between fertility rates and female labour force participation
- Sector-wise % contribution to GDP - to study if percentage contribution to GDP of sectors were influenced by changes in labour force participation rates among genders

The research questions that we have explored and studied with this data are:

1. How do the labour force participation rates along with other factors such as fertility rate, GDP and sector wise contributions affect the HDI index (score) of a country?
2. Do labour force participation rates and their difference between genders vary across continents?
3. How have the labour force participation rates of both genders changed over the years across continents?

2 *Data Analysis*

2.1 Changes in labour force participation rates of both genders changed over the years across continents

To see the how the labour force participation behaviour between genders varies across continents and how this has progressed across the years from 1990 to 2021, we have used a boxplot. As plotting over all 32 years given in the dataset would be difficult to read and interpret, we have considered the years 1990, 2005 and 2021 to study the movement of the labour force participation rates.

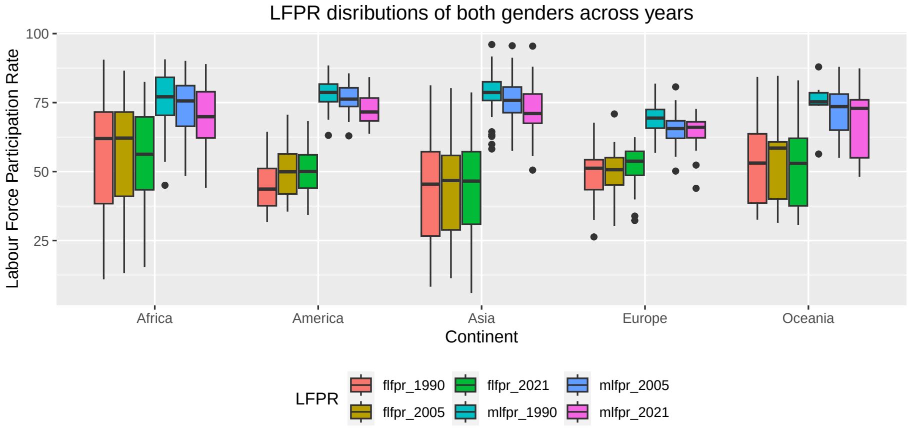


Figure 1: Boxplot of labour force participation rates across continents

In the above figure,
flfpr: female labour force particpation rate
mlfpr: male labour force participation rate

From the above plot, we can see that the median labour force participation rates for females are significantly lower than for males. Also, the spread of the participation rates are much greater for females than for males which shows that the female labour force participation rates vary greatly among countries even withing continents, whereas male labour force participation rates are spread within a significantly smaller range.

We can also see that the male median participation rate has decreased over the years across all continents, while the female participation rate has increased in all continents except in Africa where it has decreased.

2.2 HDI Rank vs behaviour of average labour force participation rates between genders

On arranging the countries in ascending order of their HDI Rank and plotting their corresponding average male and female labour force participation rates over the last five years, we get the following plot.

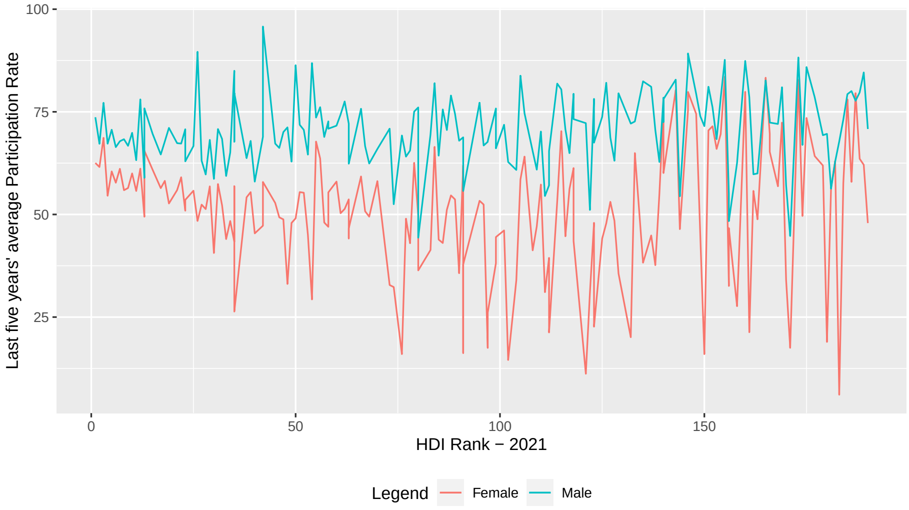


Figure 2: Difference in labour force participation rates among males and females

We can see that the countries with lower rank (i.e. better HDI) have a consistently low difference in the labour force participation rates among males and females, whereas countries with higher rank (i.e. worse HDI) have higher difference in the labour force participation rates among males and females.

2.3 Regression Analysis

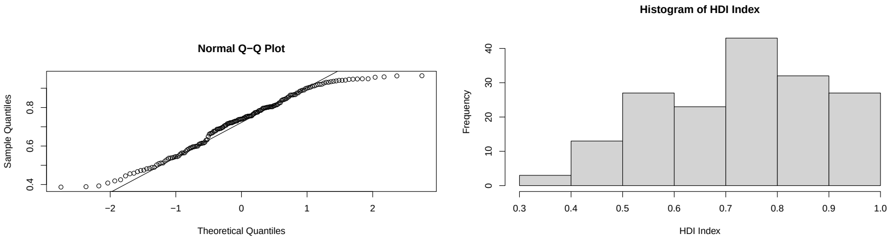
A regression analysis would be helpful in studying how the labour force participation rates along with other factors such as fertility rate, GDP and sector wise contributions affect the HDI index (score) of a country. Before proceeding with this, we first check if the dependent variable (HDI Index) is normally distributed.

Graphically, we can check if it can be assumed that the variables are normally distributed using normal QQ plots and histograms.

Normal QQ plot: Points on the normal QQ plot helps assess the univariate normality of the data. If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.

Histogram: A histogram is an informal way of checking if the data is normally distributed. This is done by constructing a histogram of the data and comparing it to a normal probability curve (bell curve). If the data were normally distributed, the distribution (the histogram) should be bell-shaped and resemble the normal distribution.

The normal QQ plot and histogram for the 'HDI Index' variable is shown below.



(a) QQ plot of HDI Index

(b) Histogram of HDI Inde

Figure 3: Graphical checks of normality

We can see from the QQ plot and histogram that the distibution of HDI Index seems to follow an approximaltely normal distibution. This indicates that a regression analysis would be an appropriate approach.

We have performed a regression analysis with the HDI index of a country as the dependent variable and the following as the independent variables or regressors:

- last five year average of female labour force participation rate (f_avg5yr)
- last five year average of male labour force participation rate (m_avg5yr)
- fertility rate (FertilityRate)
- percentage of GDP contributed by industry sector (Industry_per)
- percentage of GDP contributed by agriculture sector (Agriculture_per) and
- percentage of GDP contributed by services sector (Services_per).

The regression model thus developed has an R squared value of 0.8 which indicates a strong fit and the output of this regression is shown in Table 1.

From this regression output, we can infer the following:

- Female labour force participation rate has a positive and statistically significant coefficient, indicating that an increase in female labour force participation rate is associated with an increased HDI index.
- Fertility rate has a negative and statistically significant coefficient, indicating that an increase in fertility rate is associated with a decreased HDI

Table 1

	Dependent variable:
	HDI_Index
f_avg5yr	0.001^{**}
	(0.0004)
m_avg5yr	-0.001
	(0.001)
FertilityRate	-0.067^{***}
	(0.006)
Industry_per	0.001
	(0.001)
Agriculture_per	-0.005^{***}
	(0.002)
Services_per	0.002
	(0.001)
Constant	0.865^{***}
	(0.115)
Observations	168
R^2	0.803
Adjusted R^2	0.795
Residual Std. Error	0.069 (df = 161)
F Statistic	109.045^{***} (df = 6; 161)
Note:	*p<0.1; **p<0.05; ***p<0.01

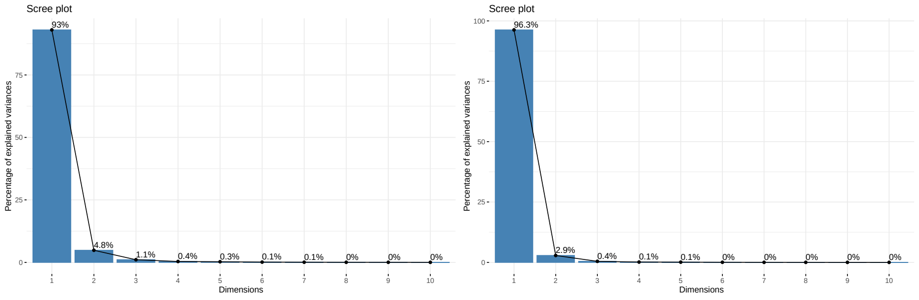
index. This is also intuitive in the sense that, when the fertility rate is high, i.e. when there are more children to take care of, female labour force participation is likely to be lesser, thus affecting the overall productivity negatively.

- Percentage contribution to GDP by the agriculture sector has a negative and statistically significant coefficient, indicating that countries where agriculture contributes majorly to the GDP are associated with a decreased HDI index. This could be because countries where agriculture is the major contributing sector are not highly developed and hence would have a lower HDI index.

2.4 Principal Component Analysis and Principal Component Regression

In our regression analysis, we have taken the average of last five years' male and female labour force participation rates. Another approach we have tried is to perform a principal component analysis (PCA) over the labour force participation rates across all 32 years (1990 to 2021) in the dataset. This was done to reduce loss of information that would occur when taking only the last five years' values and averaging them out.

On performing PCA over the labour force participation rates of males and females separately, we find that the first principal component of both captures 93% and 96% of the total variation respectively, as seen in the screeplots below.



(a) Screeplot - Male LFPR

(b) Screeplot - Female LFPR

Figure 4: PCA - Screeplots

Using the PC components instead of the 5 year average labour force participation rates in the regression model, we performed a regression analysis with the HDI index of a country as the dependent variable and the following as the independent variables or regressors: principal component of female labour force participation rate (f_pca), principal component of male labour force participation rate (m_pca), fertility rate (FertilityRate), percentage of GDP contributed by industry sector (Industry_per), percentage of GDP contributed by agriculture sector (Agriculture_per) and percentage of GDP contributed by services sector (Services_per).

The regression model thus developed has an R squared value of 0.8 which indicates a strong fit. The output of this regression is shown in Table 2.

We find that this regression does not offer any incremental benefit in terms of performance compared to the previous model and gives similar insights. We attribute this to the fact that the change in labour force participation rates for a country do not vary much year on year and hence, taking the average over the last five years' data captures almost the same amount of information as the principal component over 32 years' data.

Table 2

	Dependent variable:
	HDI_Index
f_pca	0.002^{*}
	(0.001)
m_pca	0.002^{**}
	(0.001)
FertilityRate	-0.067^{***}
	(0.006)
Industry_per	0.001
	(0.001)
Agriculture_per	-0.004^{***}
	(0.002)
Services_per	0.002
	(0.001)
Constant	0.817^{***}
	(0.117)
Observations	168
R^2	0.804
Adjusted R^2	0.796
Residual Std. Error	0.069 (df = 161)
F Statistic	109.792^{***} (df = 6; 161)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.5 Cluster Analysis of Countries based on GDP and HDI Rank

The KMeans clustering algorithm was applied to analyze the dataset containing information on GDP and HDI rank for various countries. Three distinct clusters were identified based on the similarities in these socio-economic indicators. The clusters were visualized using a scatter plot, where each cluster was represented by a different color.

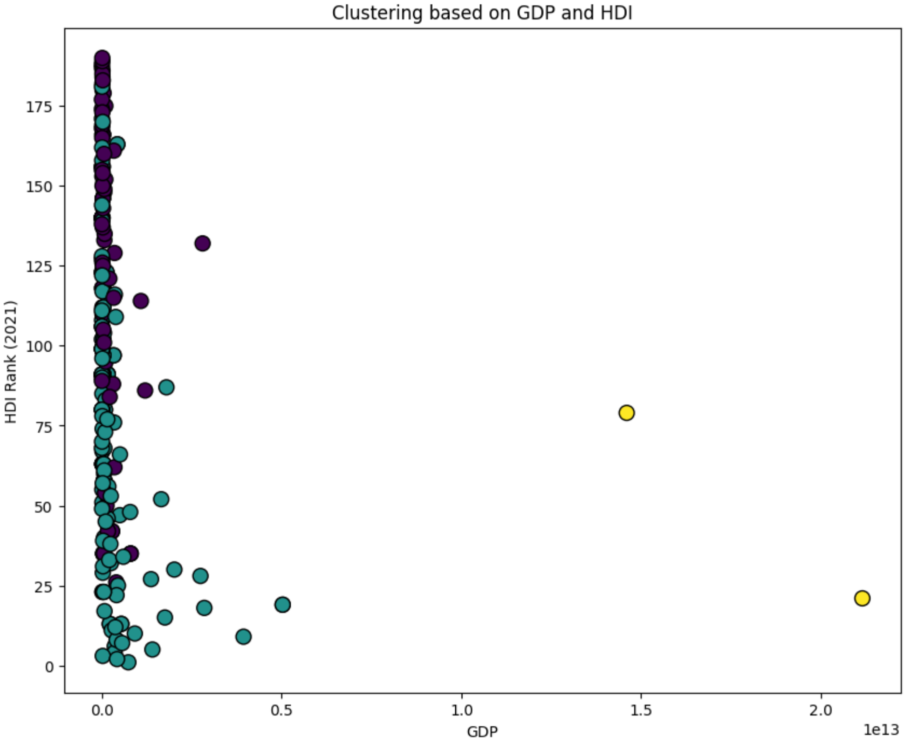


Figure 5: K-means Clustering of GDP and HDI

Cluster Interpretation:

1. **Cluster 1 (Green):** This cluster comprises countries with high GDP and high HDI rank, such as Norway, Switzerland, and Australia. These countries are characterized by strong economic performance and high levels of human development.

2. **Cluster 2 (Yellow):** Countries in this cluster exhibit low GDP and low HDI rank, including Niger, Burundi, and the Central African Republic. These nations face significant socio-economic challenges and may require targeted interventions to improve living standards and economic opportu-

nities.

3. **Cluster 3 (Purple):** This cluster encompasses countries with medium GDP and medium HDI rank, such as China, India, and Brazil. These countries occupy an intermediate position in terms of economic development and human well-being, with opportunities for further growth and advancement.

Implications and Insights:

The findings of the KMeans clustering analysis suggest a positive correlation between GDP and HDI rank, consistent with previous research. The identification of distinct clusters provides valuable insights into the socio-economic landscape of different countries and can be used to inform policy-making and development strategies.

3 *References*

- UNDP Website - Human Development Index
- Fertility rate(data from World Bank)
- GDP data from World Bank)
- Industry sector contribution to GDP data(from World Bank)
- Agriculture sector contribution to GDP data(from World Bank)
- Services sector contribution to GDP data(from World Bank)
- Link for the clustering code and output findings: Python code for clustering