Advanced data visualization

Experiment-2

Name	Parth Gandhi
UID	2021300033
Batch	Batch H
Department	COMPS A

Aim: Create advanced charts using Python on socio economic dataset. Create WordChart, Boxplot, whiskers plot Regression plot.

Description of dataset

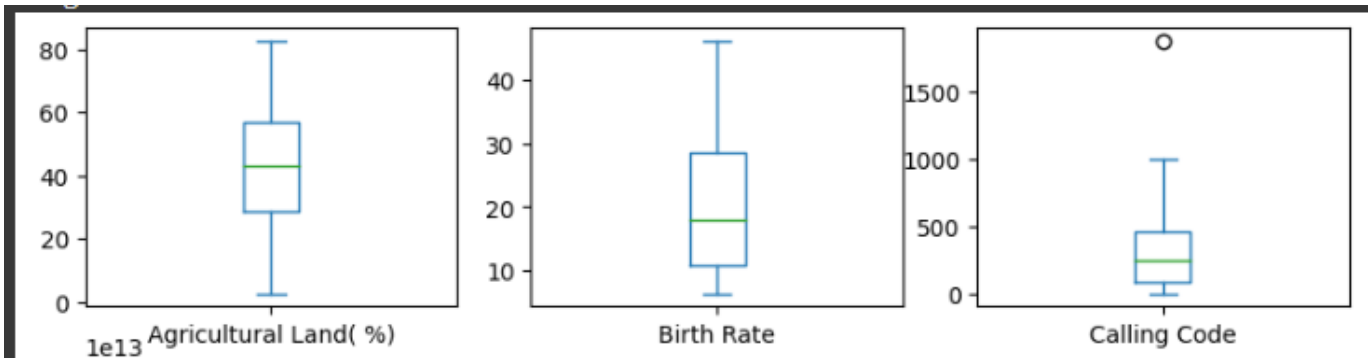
The dataset being used is a world dataset containing basic information about various countries. The dataset contains following attributes:

1. **Country**: Name of the country.
2. **Density (P/Km2)**: Population density measured in persons per square kilometer.
3. **Abbreviation**: Abbreviation or code representing the country.
4. **Agricultural Land (%)**: Percentage of land area used for agricultural purposes.
5. **Land Area (Km2)**: Total land area of the country in square kilometers.
6. **Armed Forces Size**: Size of the armed forces in the country.
7. **Birth Rate**: Number of births per 1,000 population per year.
8. **Calling Code**: International calling code for the country.
9. **Capital/Major City**: Name of the capital or major city.
10. **CO2 Emissions**: Carbon dioxide emissions in tons.
11. **CPI**: Consumer Price Index, a measure of inflation and purchasing power.
12. **CPI Change (%)**: Percentage change in the Consumer Price Index compared to the previous year.
13. **Currency Code**: Currency code used in the country.
14. **Fertility Rate**: Average number of children born to a woman during her lifetime.
15. **Forested Area (%)**: Percentage of land area covered by forests.
16. **Gasoline Price**: Price of gasoline per liter in local currency.
17. **GDP**: Gross Domestic Product, the total value of goods and services produced in the country.
18. **Gross Primary Education Enrollment (%)**: Gross enrollment ratio for primary education.
19. **Gross Tertiary Education Enrollment (%)**: Gross enrollment ratio for tertiary education.
20. **Infant Mortality**: Number of deaths per 1,000 live births before reaching one year of age.
21. **Largest City**: Name of the country's largest city.
22. **Life Expectancy**: Average number of years a newborn is expected to live.
23. **Maternal Mortality Ratio**: Number of maternal deaths per 100,000 live births.
24. **Minimum Wage**: Minimum wage level in local currency.
25. **Official Language**: Official language(s) spoken in the country.
26. **Out of Pocket Health Expenditure (%)**: Percentage of total health expenditure paid out-of-pocket by individuals.
27. **Physicians per Thousand**: Number of physicians per thousand people.
28. **Population**: Total population of the country.

29. **Population: Labor Force Participation (%)**: Percentage of the population that is part of the labor force.
30. **Tax Revenue (%)**: Tax revenue as a percentage of GDP.
31. **Total Tax Rate**: Overall tax burden as a percentage of commercial profits.
32. **Unemployment Rate**: Percentage of the labor force that is unemployed.
33. **Urban Population**: Percentage of the population living in urban areas.
34. **Latitude**: Latitude coordinate of the country's location.
35. **Longitude**: Longitude coordinate of the country's location.

Analysis of charts

1. Box Plot



1. Agricultural Land (%)

- The box plot for Agricultural Land (%) shows that the data is relatively spread out, with the interquartile range (IQR) covering values approximately from 30% to 60%.
- The median value is slightly above 40%, indicating that half of the countries have more than 40% of their land used for agriculture.
- The minimum value is around 0% and the maximum value is around 80%.
- There are no visible outliers in this distribution.

2. Birth Rate

- The Birth Rate distribution is also fairly spread out, with the IQR ranging from around 15 to 35 births per 1,000 people.
- The median birth rate is around 22, which means that half of the countries have a birth rate above this value.
- The birth rate varies from around 10 to 45 births per 1,000 people.
- Similar to Agricultural Land (%), there are no visible outliers in this plot.

3. Calling Code

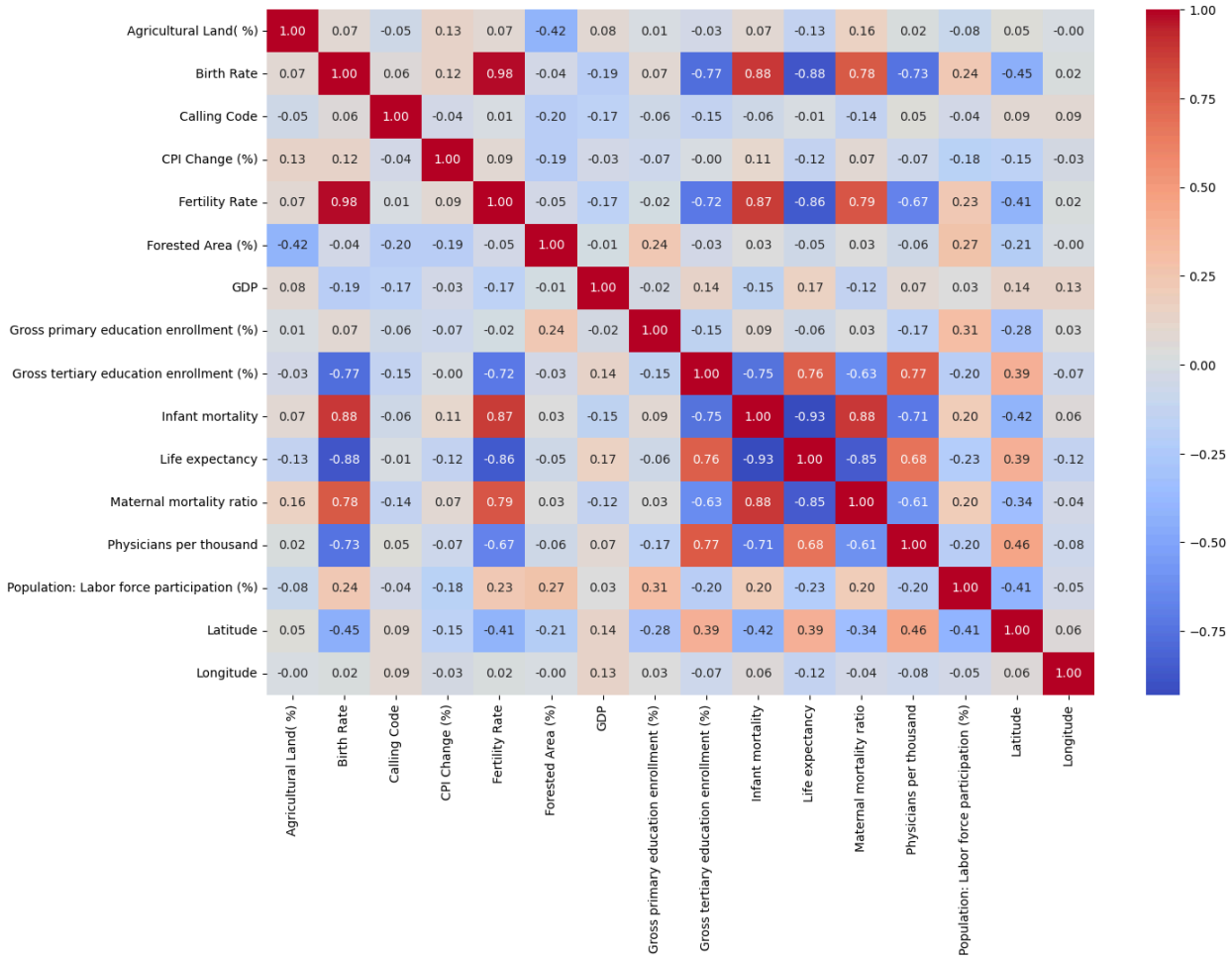
- The Calling Code plot shows a more concentrated distribution with a small IQR.
- The median calling code is quite low, indicating that most countries have smaller calling codes.
- Most of the data falls between 0 and 500, but the presence of a significant outlier is indicated by the value above 1500.
- There is a clear outlier above the whisker, which could be a country with a unique or unusual calling code (e.g., a country with a very high numeric code).

2. Word cloud

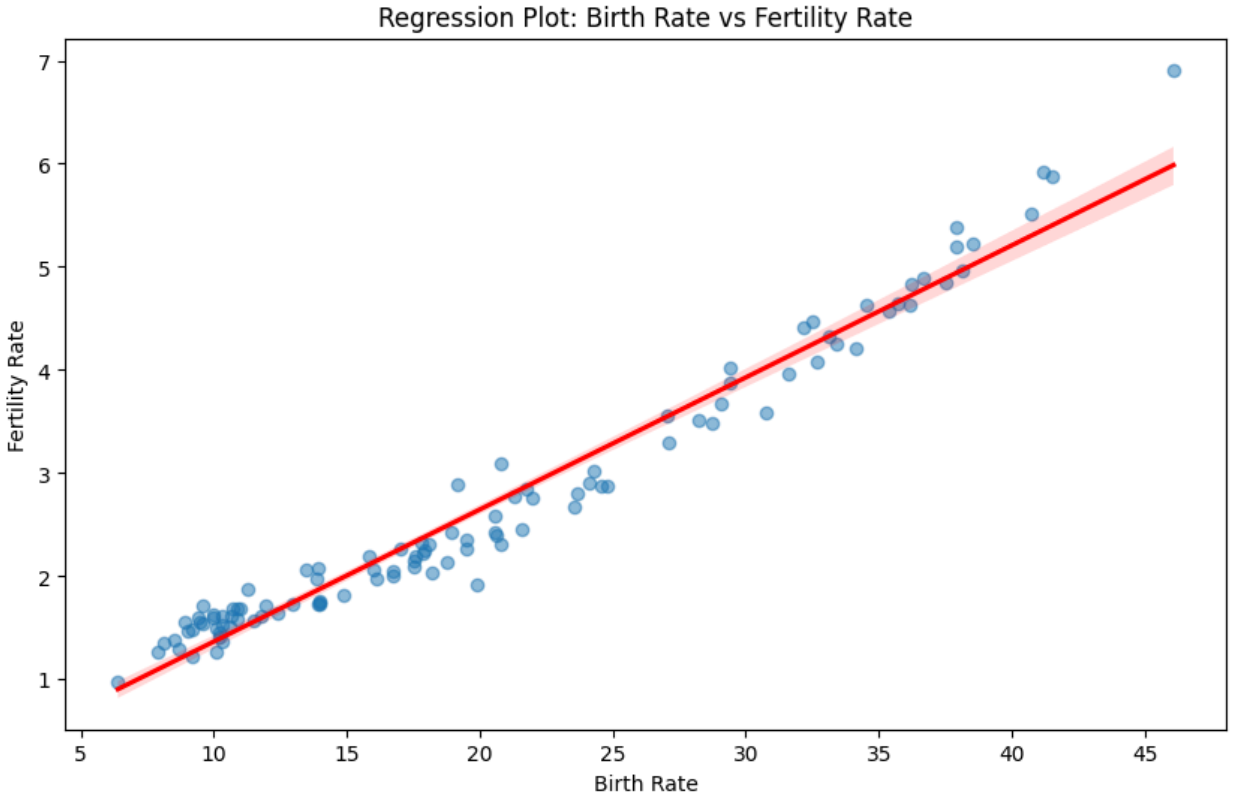
The word cloud represents the Official Languages of various countries.

1. **Dominant Languages**:
   - **English**: The largest word in the cloud, indicating that English is the most common official language among the countries in the dataset.
   - **French**: Another prominently featured language, suggesting it is also widely spoken as an official language in many countries.
   - **Spanish, Arabic, Portuguese**: These languages are also fairly large, indicating their widespread use as official languages.
2. **Moderately Prominent Languages**:
   - **Russian, Chinese, German, Hindi**: These languages are moderately sized in the word cloud, reflecting their official status in several countries, but less so than the top few.
3. **Less Common Languages**:
   - **Swahili, Albanian, Vietnamese, Bengali**: These languages are smaller in the word cloud, suggesting they are official languages in fewer countries.
4. **Rare Official Languages**:
   - **Maltese, Burmese, Slovak, Lao**: These languages are among the smallest, indicating they are official in very few countries.

3. Regression Plot



A correlation heatmap was first plotted to see the similarities between the attributes. As observed, the correlation between birth rate and fertility rate was highest, hence these attributes were chosen for the regression plot.



- The scatter plot and the fitted regression line indicate a strong positive correlation between the birth rate and fertility rate. As the birth rate increases, the fertility rate also tends to increase.
- The data points are closely aligned with the regression line, suggesting that a linear model is a good fit for this data. The line is drawn in red, with a shaded area representing the confidence interval of the regression.
- There is at least one clear outlier on the upper right side of the plot, where both the birth rate and fertility rate are significantly higher compared to the rest of the data points.
- The data points appear more densely packed at lower birth rates (between 5 and 20) and fertility rates (between 1 and 3). As the birth rate increases, the spread of the fertility rate also increases slightly, but the general trend remains strong and linear.
- The shaded area around the regression line is relatively narrow, which indicates that the model's predictions are fairly confident. However, this interval widens slightly at higher values, reflecting greater uncertainty.

Conclusion: In this experiment, we created and analyzed various charts, including regression plot, box plot/whisker plot, and word cloud, each revealing unique patterns and trends within the data.

The regression plot revealed clear patterns and relationships, such as the strong positive correlation between birth rate and fertility rate. The box plot provided a clear view of the distribution and variability of key metrics like Agricultural Land percentage, Birth Rate, and Calling Codes, with notable outliers identified in the Calling Code data. The word cloud offered a visual representation of the linguistic diversity across countries, emphasizing the global prevalence of languages like English, French, and Spanish.