**INFANT MORTALITY RATE PREDICTION**

**INTRODUCTION**

New-born refers to a baby from birth to about two months of age. Infant can be considered children anywhere from birth to 1 year old. Infant mortality is the death of an infant before their first birthday. The infant mortality rate is the number of infant deaths for every 1,000 live births. In addition to giving us essential information about maternal and infant health, the infant mortality rate is a crucial marker of society's overall health. According to World Health Organization (WHO), from the end of the neonatal period and through the first five years of life, the leading causes of death are pneumonia, diarrhea, congenital disabilities, and malaria. Apart from this, within the first month after birth, short gestation and low birth weight also form the leading cause of neonatal mortality, followed by congenital malformations and maternal complications.

Education in men and majorly in women plays a vital role in their decision-making capabilities concerning reproduction and healthcare. Literacy helps in population stabilization and better infant care, which is reflected by lower birth rates and infant mortality rates (IMRs). According to a study published in the Journal of Family Medicine and Primary Care (JFPMC), female literacy is highly important for population stabilization and infant health. Additionally, per capita GDP also plays an essential role in infant mortality. There is a robust relationship between per capita GDP and infant mortality: on average, a 1% decrease in per capita GDP increases infant mortality of between 0.24 and 0.40 per 1000 children born.

IMR is regarded as an essential indicator of population health. Low mortality rates ensure better healthcare. Lower birth rates are associated with less growth, a more rapidly aging population, and, hence, slower economic expansion. Therefore, it is vital to study this issue and find a solution. This study aims to see the impact of Literacy Rate and per capita GDP on infant mortality rate.

**DATA DESCRIPTION AND SUMMARY**

The data is collected from <https://data.gov.in/> from 2011. The analysis is done using SPSS and Excel. The table contains 34 observations and four columns, namely State, Mortality Rate, Literacy Rate, Per capita GDP for all the states of India. The data description is given below:

Table 1. Data Description

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| **State** | Contains state names |
| **Mortality Rate** | Contains infant mortality rate for each State |
| **Literacy Rate** | Contains literacy rate of both men and women combined for each State. |
| **Per Capita GDP** | Contains GDP of the particular State. |

**DESCRIPTION OF MODELS**

The following models are used for this study-

* Cluster Analysis- multivariate data mining technique aims to group objects (e.g., products, respondents, or other entities) based on user-selected characteristics or attributes.
* Multiple linear regression- Used to find out the regression coefficients for the factors to see their effect in terms of direction and magnitude on IMR.

**MODEL FITTING**

The analysis is done first to understand the data by visualization. Various plots are used, such as scatter/dot plots, box plots, and histograms.

DATA ANALYSIS

The correlation is calculated between all the variables. Pearson Correlation is used here since the data is numeric.

Table 2. Pearson Correlation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **MORTALITYRATE** | **LITERACYRATE** | **PERCAPITAGCP** |
| **MORTALITYRATE** | Pearson Correlation | 1 | -.705\*\* | -.525\*\* |
| Sig. (2-tailed) |  | <.001 | .002 |
| N | 34 | 34 | 33 |
| **LITERACYRATE** | Pearson Correlation | -.705\*\* | 1 | .579\*\* |
| Sig. (2-tailed) | <.001 |  | <.001 |
| N | 34 | 34 | 33 |
| **PERCAPITAGCP** | Pearson Correlation | -.525\*\* | .579\*\* | 1 |
| Sig. (2-tailed) | .002 | <.001 |  |
| N | 33 | 33 | 33 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | |

From Table 2, it can be understood that both Literacy rate and Per capita GDP are inversely correlated with Mortality rate. In this sense, if the Literacy rate is lower, the Mortality rate increases.

Figure 1 shows the box plot of the Mortality rate. The figure shows that the variable is normally distributed, falling between 31 and 187. Noticeably, there is no outlier.

img.emf

Figure 1. Box plot of Mortality rate

Figure 2 shows the box plot of the Literacy rate. The distribution is normal. The range of the variable is between 61.80 and 94.

img.emf

Figure 2. Boxplot of Literacy Rate

Figure 3 shows the box plot of Per capita GDP. It is seen that there are a lot of outliers in the variable. But, since the data differs for each State and the data is limited, the outliers cannot be eliminated.

img.emf

Figure 3. Boxplot of Per capita GDP

Figure 4 shows the data distribution for each State in this scatterplot plotted for each variable.

img.emf

Figure 4. Scatterplot for all variables to State

Figure 5 shows the frequency of each value in the Mortality rate. It is seen that most of the observations are unique. There are only three observations that are repeated throughout the data.

img.emf

Figure 5. Frequency of values in Mortality Rate

Figure 6 shows the distribution of values in the data in Mortality data. It is observed that most of the IMR values are in the range ~75, as shown in Figure 6.

img.emf

Figure 6. Distribution of data in Mortality rate

Figure 7 shows the distribution of Per Capita GDP in the data. From the plot, it can be seen that the most common value in the data is approximately equal to 50000.

img.emf

Figure 7. Distribution of Per capita GDP

In Figure 8, the scatterplot of Literacy Rate by Mortality Rate is given. From this plot, we can see that the correlation between Literacy Rate and Mortality Rate is negative.

img.emf

Figure 8. Scatter plot of Mortality Rate vs. Literacy Rate

From the scatterplot shown below in Figure 9, like Figure 8, the variables are negatively correlated. Also, there is a noticeable outlier in between 200000 and 300000.

img.emf

Figure 9. Scatterplot of Mortality Rate vs Per Capita GDP

CLUSTER ANALYSIS

After the primary analysis was done, we went ahead with cluster analysis to see the similarity in the data and detect the patterns. Two-step Clustering is used. The two clusters obtained from this analysis are the High Mortality Rate and Low Mortality Rate.

Table 3: Cluster names

|  |  |
| --- | --- |
| **Cluster** | **Cluster Name** |
| **1** | High Mortality Rate |
| **2** | Low Mortality Rate |

Table 4: Cluster sizes

|  |  |
| --- | --- |
| **Cluster** | **Cluster sizes** |
| **1** | 13 (39.4%) |
| **2** | 20 (60.6%) |

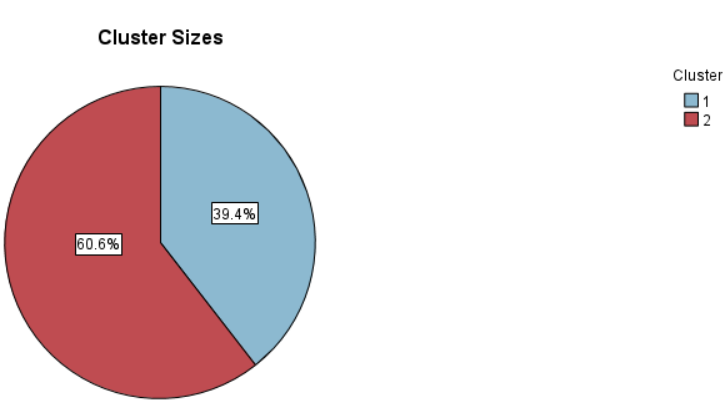
Cluster one has the highest mortality rate, with less literacy and per capita GDP. This shows that the Mortality rate is inversely proportional to the Literacy Rate and Per Capita GDP. The cluster sizes are visualized below in Figure 10.

Figure 10: Cluster Sizes

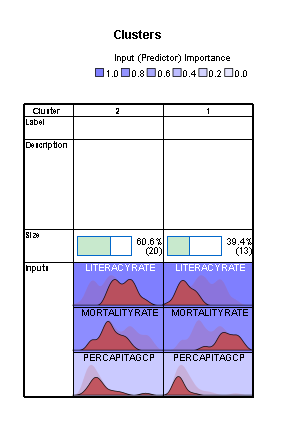


Figure 11: Cluster Table

The cluster table (Figure 11) shows the detailed view of the clusters. The variable State is not included in the clustering, as it has zero importance in the process. Also, the second cluster is the highest, with 60.6% values present in it. According to the clustering results shown in Figures 11 and 12, Literacy Rate is the essential variable in the analysis, followed by Mortality Rate and Per Capita GDP.

Figures 13, 14, and 15 show the values distribution of Mortality Rate, Per Capita GDP, and Literacy Rate in Cluster 1, respectively.

Figures 16, 17, and 18 show the values distribution of Mortality Rate, Per Capita GDP, and Literacy Rate in Cluster 2.

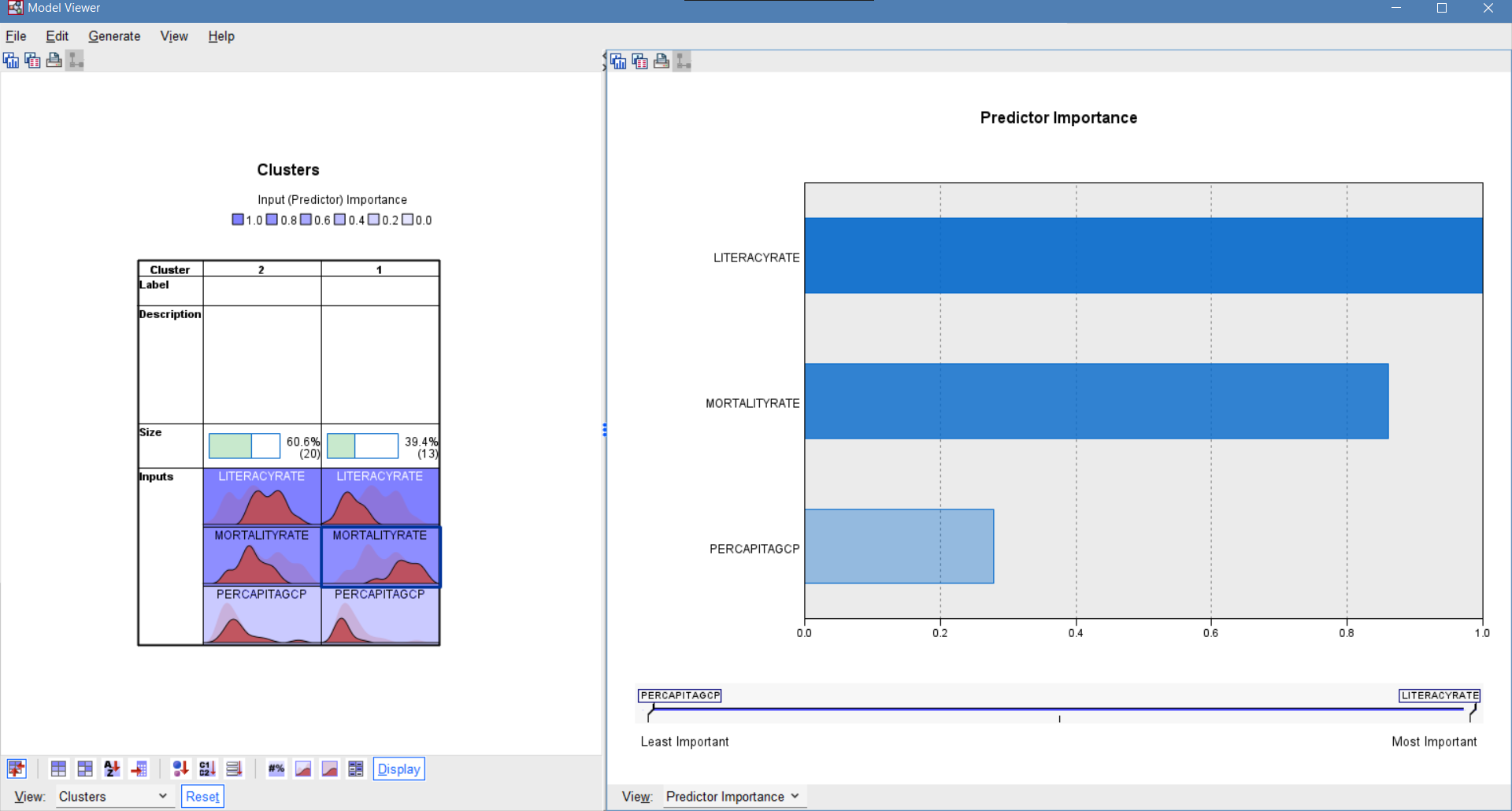


Figure 12: Predictor Importance

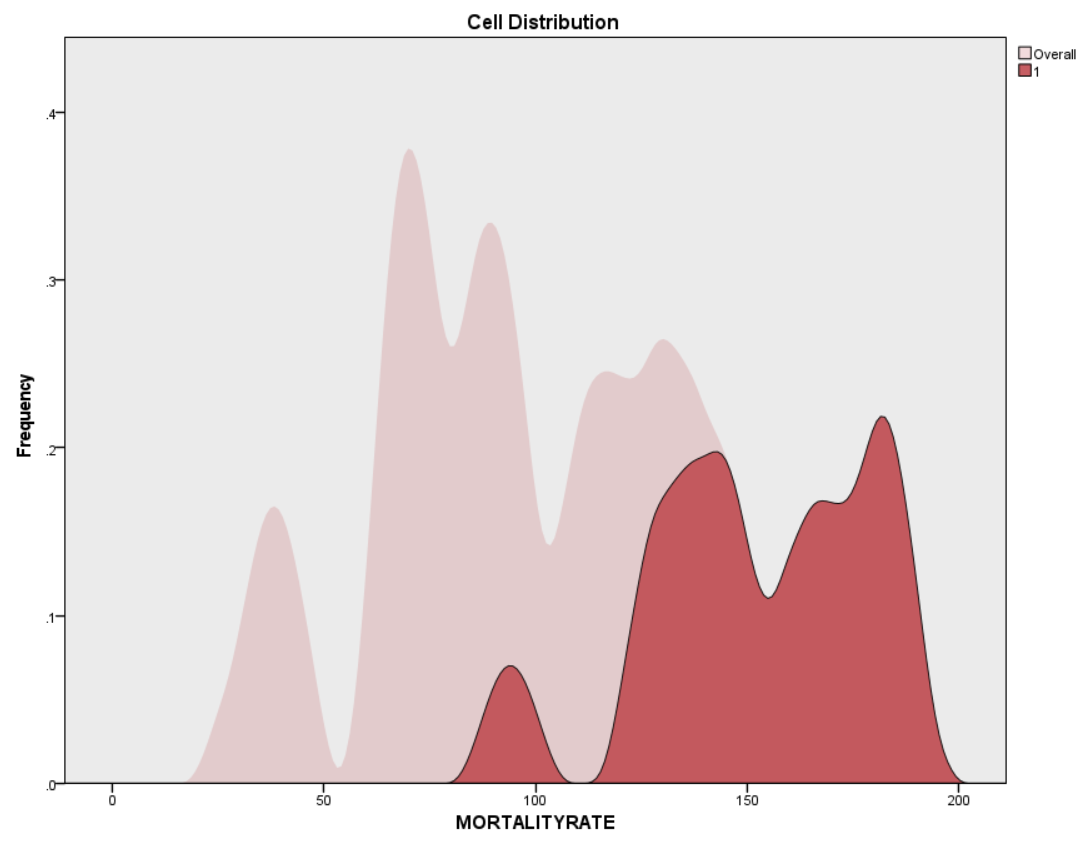


Figure 13: Data distribution of Mortality rate in Cluster 1

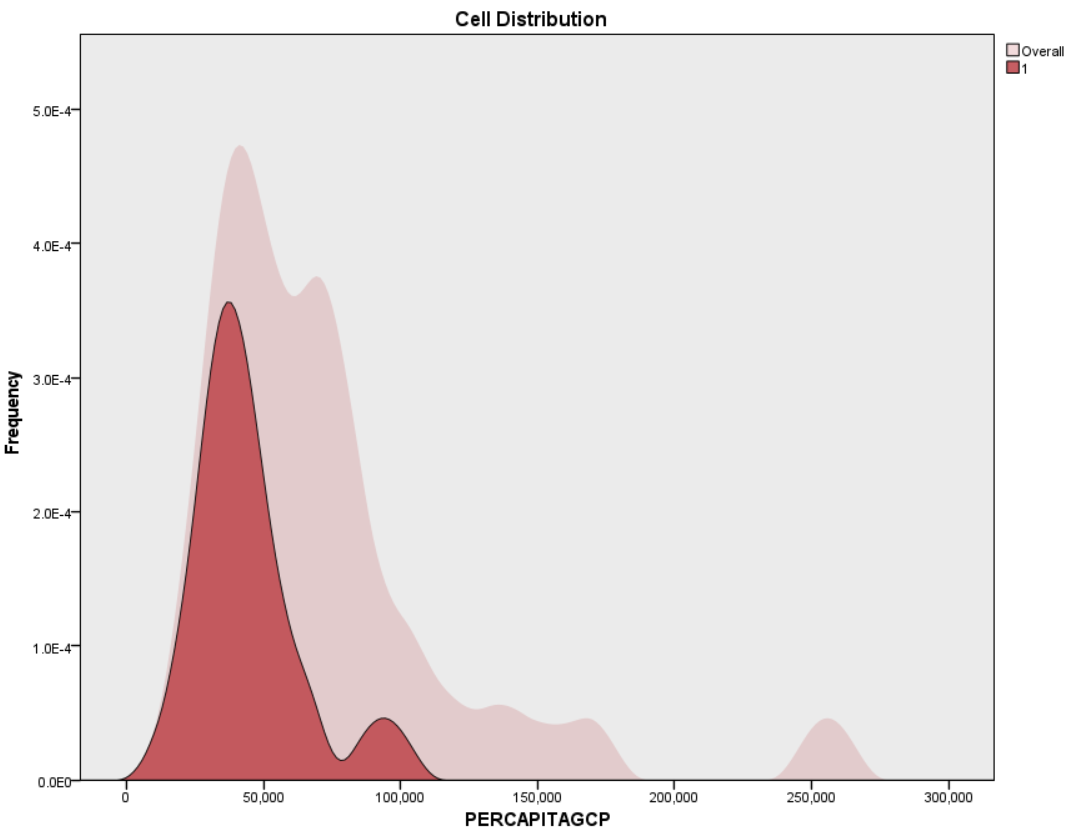


Figure 14: Data distribution of Per capita GDP in Cluster 1

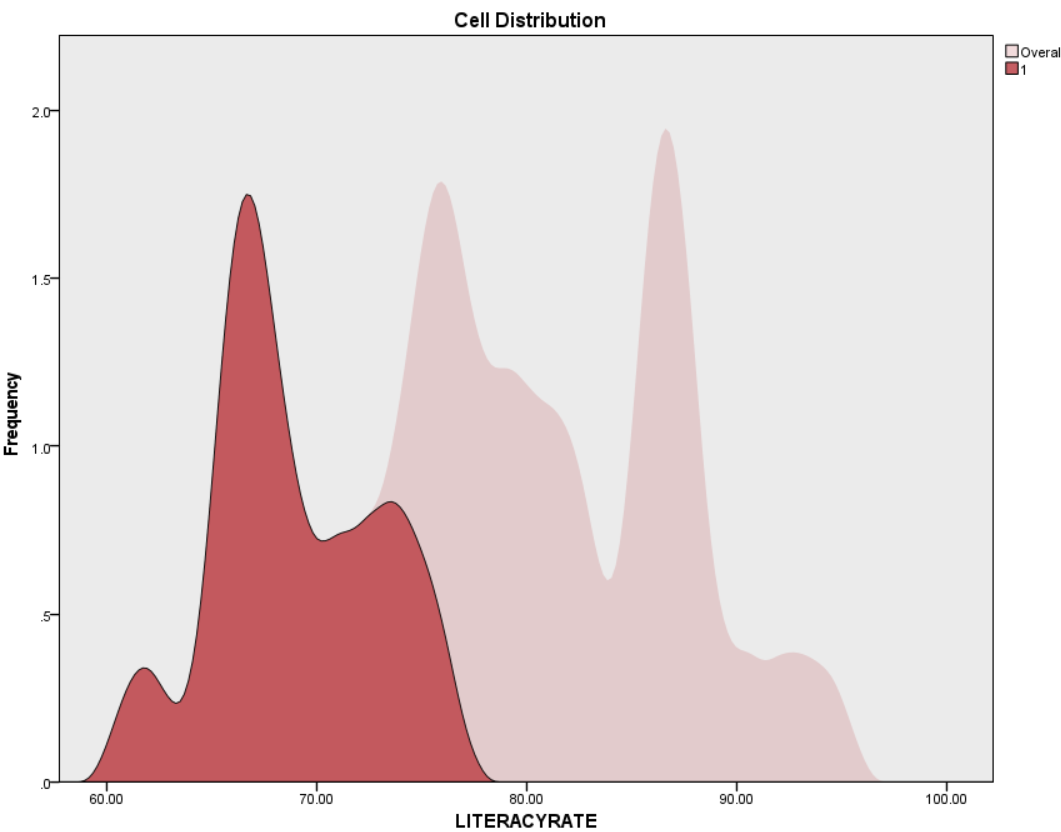


Figure 15: Data distribution of Literacy rate in Cluster 1

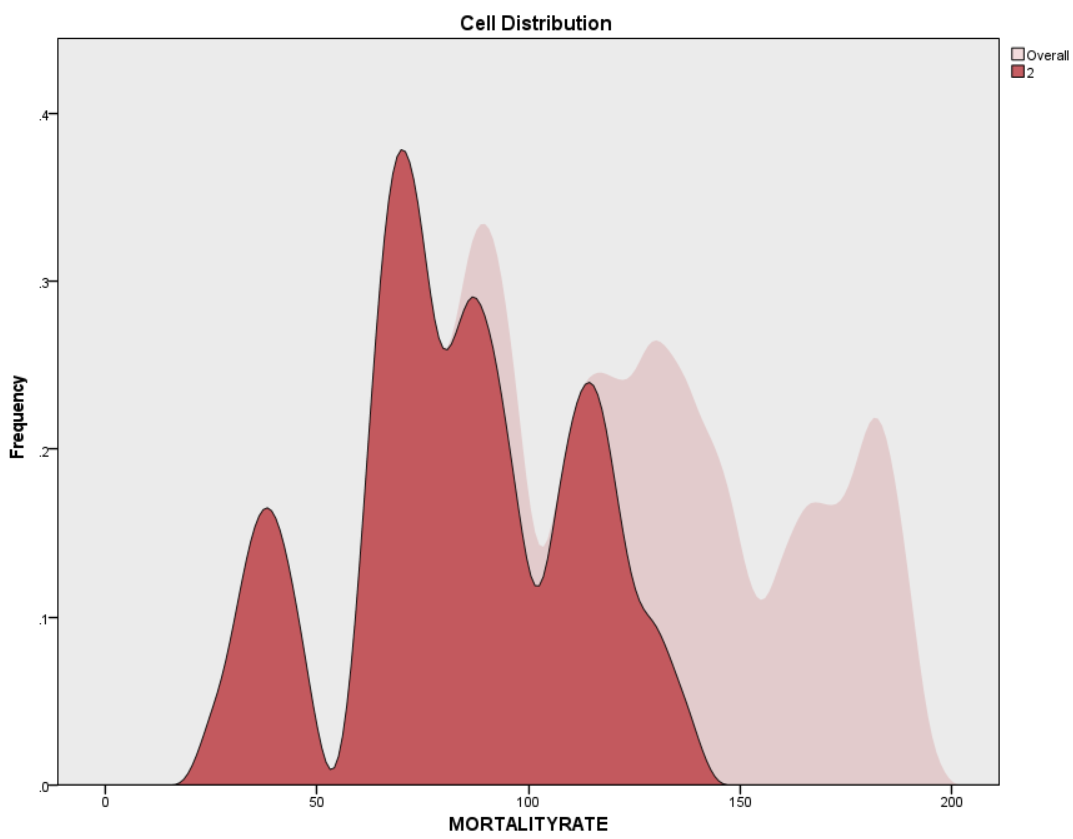


Figure 16: Data distribution of Mortality Rate in Cluster 2

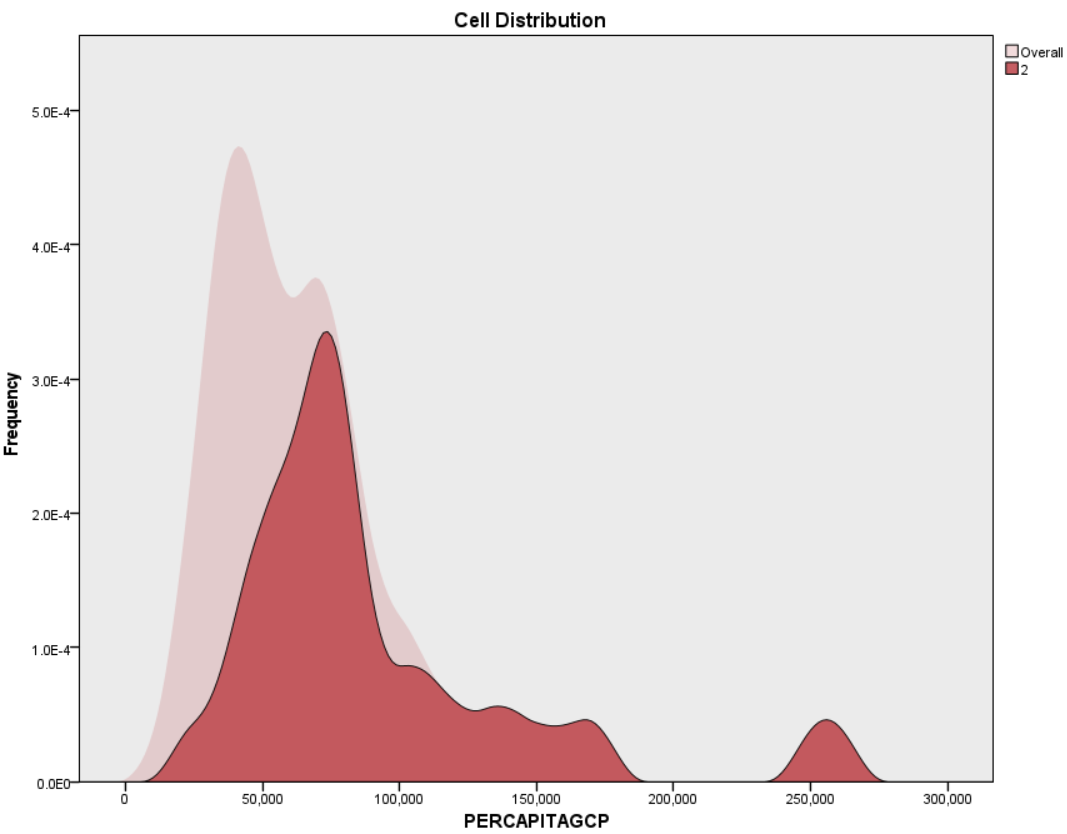


Figure 17: Data distribution of Per Capita GDP in Cluster 2

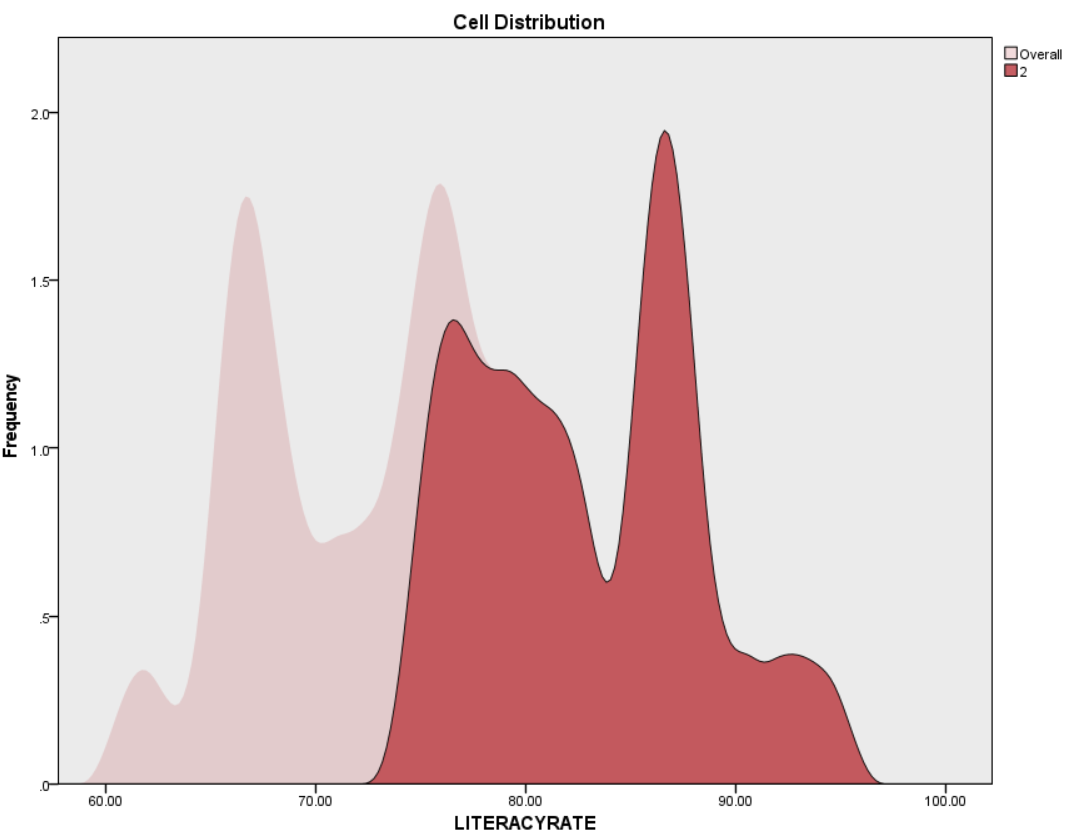


Figure 18: Data Distribution of Literacy Rate in Cluster 2

Silhouette measure of Cohesion and Separation is used to check the cluster quality. As shown in Figure 19, the cluster quality is above 0.5, which is reliable.

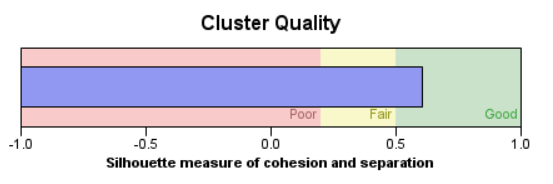


Figure 19: Cluster quality

LINEAR REGRESSION

State, Literacy Rate, and Per Capita GDP are the independent variables, and Mortality Rate is the dependent variable. The independent variables are taken into the regression function to predict the dependent variable, Mortality Rate. The descriptive statistics of the variables are given in Table 5.

Table 5: Descriptive Statistics

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mean** | **Std. Deviation** | **N** |
| **MORTALITYRATE** | 110.45 | 43.755 | 33 |
| **LITERACYRATE** | 77.4785 | 8.49191 | 33 |
| **PERCAPITAGCP** | 71755.09 | 47616.143 | 33 |

The mean and standard deviation for all the variables is given in table 5.

The correlation table is given in Table 6. This table shows that the correlation between Mortality Rate and Literacy Rate, Per Capita GDP is negative, as seen in the analysis.

Table 6: Correlations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **MORTALITYRATE** | **LITERACYRATE** | **PERCAPITAGCP** |
| **Pearson Correlation** | MORTALITYRATE | 1.000 | -.701 | -.525 |
| LITERACYRATE | -.701 | 1.000 | .579 |
| PERCAPITAGCP | -.525 | .579 | 1.000 |
| **Sig. (1-tailed)** | MORTALITYRATE | . | <.001 | <.001 |
| LITERACYRATE | .000 | . | .000 |
| PERCAPITAGCP | .001 | .000 | . |
| **N** | MORTALITYRATE | 33 | 33 | 33 |
| LITERACYRATE | 33 | 33 | 33 |
| PERCAPITAGCP | 33 | 33 | 33 |

Model Summary is given in Table 7. R-value is 0.716, which shows the strong linear relationship between the variables. An adjusted R square of 0.480 indicates the variability in the outcome is 48%. Since the data size is small, given the data is based on every State in India, the R square value is comparatively large. But as the sample size increases, the adjusted R square significantly decreases. The Std. Error of the estimate determines the variability surrounding a coefficient estimate. A standard error of 31.545 indicates the data is more spread out and is not clustered about the mean. The standard error will be less if the data is surrounded by the mean Mortality rate, which is 110.45 (table 5).

Table 7: Model Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R** | **R Square** | **Adjusted R Square** | **Std. Error of the Estimate** |
| **1** | .716a | .513 | .480 | 31.545 |

1. Predictors: (Constant), PERCAPITAGCP, LITERACYRATE
2. Dependent Variable: MORTALITYRATE

ANOVA table is given in Table 8. This table lets us know whether the regression model is statistically significant. The significance <0.001 indicates that the model is highly significant.

Table 8: ANOVA

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **Sum of Squares** | **df** | **Mean Square** | **F** | **Sig.** |
| **1** | Regression | 31410.674 | 2 | 15705.337 | 15.782 | <.001b |
| Residual | 29853.508 | 30 | 995.117 |  |  |
| Total | 61264.182 | 32 |  |  |  |

1. Dependent Variable: MORTALITYRATE
2. Predictors: (Constant), PERCAPITAGCP, LITERACYRATE

Coefficients with their significance level are given in table 9 and 10. Since the data is inversely correlated, if the Mortality rate increases by 1 unit, the Literacy rate should be decreased by -3.078. The intercept value (Constant) in the table shows that if Literacy Rate is 0, the Mortality rate would be 360.745. Also, the significance of Per Capita GDP is not so reliable compared to the significance of Literacy Rate.

Table 9. Model Coefficients with Standardized and Unstandardized Values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **Unstandardized Coefficients** | | **Standardized Coefficients** | **t** | **Sig.** |
| **B** | **Std. Error** | **Beta** |
| **1** | (Constant) | 360.745 | 57.328 |  | 6.293 | <.001 |
| LITERACYRATE | -3.078 | .806 | -.597 | -3.821 | <.001 |
| PERCAPITAGCP | .000 | .000 | -.179 | -1.144 | .262 |

1. Dependent Variable: MORTALITYRATE

Table 10: Model Coefficients with 95% Confidence Interval

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | | **95.0% Confidence Interval for B** | |
| **Lower Bound** | **Upper Bound** |
| **1** | (Constant) | 243.666 | 477.824 |
| LITERACYRATE | -4.723 | -1.433 |
| PERCAPITAGCP | .000 | .000 |

1. Dependent Variable: MORTALITYRATE

Residual plot is shown in Figure 20. The results follow the normal distribution.

img.emf

Figure 20. Residual Plot for Regression Model

The normal P-P plot of standardized residual is given in Figure 21. The values are hovering approximately over the regression line.

img.emf

Figure 21. Normal P-P Plot of Regression Standardized Residual

The Scatter plot for Predicted vs. Residual is shown in Figure 22.

img.emf

Figure 22. Predicted value vs. Standardized Residual

CONCLUSION

The infant mortality rate measures human infant deaths in a group younger than one year of age. It is an important indicator of the overall physical health of a community. Preserving the lives of newborns has been a long-standing issue in public health, social policy, and humanitarian endeavors. High infant mortality rates are generally indicative of unmet human health needs in sanitation, medical care, nutrition, and education. The infant mortality rate is an age-specific ratio used by epidemiologists, demographers, physicians, and social scientists to better understand the extent and causes of infant deaths. To compute a given year's infant mortality rate in a particular area, one would need to know how many babies were born alive in the area and how many babies who were born alive died before their first birthday during that time.

The literacy rate influences IMR significantly. Educating men and women essentially reduced IMR. Female literacy plays a critical role in society's overall growth and development. Research proved that children taken care of by literate mothers get all-round development in every aspect of their lives. Health and income are strongly correlated across the countries, yet the extent to which improvements in income have a causal effect on health status remains controversial.

In this study, we tried to determine the relationship between infant mortality and literacy rate, per capita GDP in India from 2011. From the analysis, we were able to find out that both the factors are negatively correlated with IMR. From the findings, we have found that the literacy rate has highly negative effects on infant mortality, showing that the high literacy rate leads to a lower rate of infant mortality. Though GDP has a lower significance than the literacy rate, it is still negatively related to the mortality rate, similar to the literacy rate. The highest GDP corresponds to lower mortality.

So, to maintain low IMR, it is essential to educate the population of India. Besides, it is also necessary to take good care of the infants, take good care of oneself during pregnancy, take care of birth defects at the earliest, handle maternal pregnancy complications, etc.

REFERENCES

[1] <https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm#:~:text=Infant%20mortality%20is%20the%20death,for%20every%201%2C000%20live%20births>

[2] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4649870/#:~:text=IMR%20of%20the%20states%20were,related%20(P%20%3D%200.630)>.

[3] <https://openknowledge.worldbank.org/bitstream/handle/10986/4916/rest_a_00084.pdf;sequence=1>

[4] <https://www.sciencedirect.com/science/article/pii/S1877042813039128>

[5] <https://data.gov.in/>

[6] <https://data.gov.in/catalog/state-wise-infant-mortality-rate>

[7] <https://data.gov.in/resource/literacy-rate-india-nsso-and-rgi>

[8] <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026856>

[9] <https://www.britannica.com/science/infant-mortality-rate>

[10] <https://link.springer.com/article/10.1007/s10708-019-10041-1>