CENSUS DATA ANALYSIS OF ERNAKULAM DISTRICT KERALA - 2011

Dissertation submitted to

ST. ALBERT’S COLLEGE (AUTONOMOUS), ERNAKULAM

(AFFILIATED TO MAHATMA GANDHI UNIVERSITY)

In partial fulfillment of the requirement

For the award of degree of

INTEGRATED M.Sc. PROGRAMME IN BASIC SCIENCES –STATISTICS

Submitted by

ANANT M S

Reg. No. 2003730001

Under the guidance of

Ms. Suhana E N

REV. DR. A. O. KONNULLUY MEMORIAL RESEARCH CENTER

AND DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. ALBERT’S COLLEGE(AUTONOMOUS)

ERNAKULAM,KERALA,INDIA

APRIL 2025

CENSUS DATA ANALYSIS OF ERNAKULAM DISTRICT KERALA - 2011

Dissertation submitted to

ST. ALBERT'S COLLEGE (AUTONOMOUS), ERNAKULAM

(AFFILIATED TO MAHATMA GANDHI UNIVERSITY)

In partial fulfillment of the requirement

For the award of degree of

INTEGRATED M.Sc. PROGRAMME IN BASIC SCIENCES –STATISTICS

Submitted by

ANANT M S

Reg. No. 2003730001

Under the guidance of

Ms. Suhana E N

REV. DR. A. O. KONNULLUY MEMORIAL RESEARCH CENTER

AND DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. ALBERT'S COLLEGE(AUTONOMOUS)

ERNAKULAM,KERALA,INDIA

APRIL 2025

ST. ALBERT'S COLLEGE (AUTONOMOUS), ERNAKULAM

(AFFILIATED TO MAHATMA GANDHI UNIVERSITY)

DEPARTMENT OF MATHEMATICS AND STATISTICS

CERTIFICATE

Certified that this dissertation '**CENSUS DATA ANALYSIS OF ERNAKULAM DISTRICT KERALA – 2011**' is the bonafied work of **ANANT MS** who carried out the project under my supervision.

Ms.Suhana E N

Department of Mathematics and Statistics

St. Albert's College(Autonomous)

Ernakulam

Counter signed by,

Dr. Divya Mary Daise S

Head of the Department

St. Albert's College(Autonomous),Ernakulam

Submitted for the project viva-voice examination held on ………………

INTRNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

I,Anant MS hereby declare that this dissertation entitled ‘**CENSUS DATA ANALYSIS OF ERNAKULAM DISTRICT KERALA – 2011’** is an authentic record of the original work done by me under the guidance of **Ms. Suhana E N**, Department of Mathematics and Statistics ,St. Albert’s College (Autonomous), Ernakulam

I also declare that this dissertation has not been submitted by me fully or partially for the awards of any degree , diploma , title or recognition earlier.

ANANT M S

10^{TH} Semester Integrated M.Sc. Programme In Basic Sciences-Statistics

St. Albert’s College(Autonomous)

Ernakulam

Date:

ACKNOWLEDGEMENT

I would like to take this opportunity to thank one and all who provided valuable advice and assistance, without which this project would not have been successful.

I am extremely grateful to **Dr. Joseph Justin Rebello**, Principal of St. Albert's College (Autonomous), Ernakulam.

I express my sincere gratitude to our teachers for their valuable advice towards the completion of my project work ,especially to **Dr. Divya Mary Daise S**, HoD, Department of Mathematics and Statistics and **Ms. Suhana E N** , my project coordinator and guide, Department of Mathematics and Statistics , St. Albert's College(Autonomous), Ernakulam.

I am also thankful to all the staff of the library, St. Albert's College(Autonomous), Ernakulam, for their valuable support and cooperation.

I am also deeply indebted to **Dr. T. M. Jacob**, HoD, Department of Statistics, Nirmala College, Muvattupuzha, for his insightful guidance and support throughout this project.

My sincere thanks go to the **CHAIRMAN**, **Greater Cochin Development Authority (GCDA)** for their cooperation and valuable inputs, which enriched the project. I am particularly grateful to **Subash S**, Town Planner, GCDA, for his expert guidance and assistance, which significantly contributed to the project's success.

ANANT MS

10TH Semester Integrated M.Sc. Programme In Basic Sciences-Statistics

St. Albert's College(Autonomous)

CONTENTS

Chapter 1: Introduction

1.1 Background

Census data serves as a vital resource for understanding the demographic, socioeconomic, and cultural characteristics of a region. This study focuses on analyzing the 2011 Census data of Ernakulam district, Kerala, to uncover meaningful insights into various indicators such as literacy rates, sex ratios, working populations, and other socioeconomic factors. Ernakulam, being a prominent district in Kerala, is known for its economic significance and diverse population. The district includes both rural and urban areas, offering a balanced representation of the socioeconomic dynamics of the state.

Kerala, often referred to as "God's Own Country," has been recognized globally for its achievements in human development. Its unique model of development emphasizes high social indicators, particularly in education, health, and gender equality, despite its relatively modest economic performance. Within this context, Ernakulam plays a pivotal role as an epicenter of commerce, tourism, and cultural heritage. It is home to Kochi, the largest urban agglomeration in Kerala, which acts as a gateway to the state's economy through its major port and vibrant industries.

Demographic studies in regions like Ernakulam are crucial for multiple reasons. First, they help us comprehend the dynamic interplay between urbanization and socioeconomic progress. Second, they enable the identification of regional disparities, such as those between rural and urban areas, in access to resources, education, and employment opportunities. Third, such studies are integral for understanding population structures and transitions, which are essential for planning sustainable development strategies.

The 2011 Census provides a snapshot of Ernakulam during a significant period of transition, where modernization, globalization, and migration were shaping its demographic and economic landscape. By analyzing this data, the study seeks to illuminate patterns and trends that have implications for regional development policies. Additionally, insights into factors like work participation rates, literacy levels, and gender disparities offer valuable inputs for addressing challenges and leveraging opportunities in the district.

This research is not only a statistical exercise but also a narrative that connects numbers with the lived realities of the district's population. It highlights the diversity and complexity of Ernakulam, from its bustling urban centers to its tranquil rural hinterlands. The findings of this study aim to serve as a foundation for policymakers, academics, and other stakeholders who are working towards inclusive and sustainable development in Kerala.

1.2 Importance of the Study

Understanding the demographic and socioeconomic structure of Ernakulam district through a detailed analysis of census data is essential for several reasons:

1. **Policy Formulation:** The findings can guide policymakers in designing effective and targeted interventions for socioeconomic development.
2. **Resource Allocation:** Insights into key indicators like literacy rates and work participation can help allocate resources more efficiently.
3. **Monitoring Progress:** The analysis provides a baseline for comparing future census data and tracking progress in key areas like education, gender equality, and workforce participation.
4. **Identifying Disparities:** By comparing rural and urban areas, the study highlights disparities and suggests areas for improvement.
5. **Academic Contribution:** The study contributes to the growing body of literature on demographic analysis, particularly in the Indian context.

In a rapidly changing economic and social environment, this study provides a snapshot of Ernakulam's demographics in 2011, offering a foundation for further research and development planning.

1.3 Objectives of the Chapter

The purpose of this chapter is to provide an introduction to the study, establish its importance, and outline the structure of the report. The subsequent sections of this chapter will delve into the broader significance of the study and its context within Kerala’s demographic landscape.

1.4 Structure of the Study

This report is organized into five chapters:

1. **Introduction** – Overview of the study and its significance.
2. **Objectives and Data Description** – Outlining the research objectives and dataset specifics.
3. **Methodology** – Details of the analytical approaches and statistical techniques used.
4. **Data Analysis** – Comprehensive analysis and interpretation of the findings.
5. **Conclusion** – Summary of key findings and recommendations.

Chapter 2: Objectives and Data Description

2.1 Objectives of the Study

The primary objectives of this study are:

1. To analyze demographic indicators such as sex ratio and literacy rates.
2. To examine the working and non-working populations across different regions.
3. To explore the relationships between socioeconomic indicators like literacy and work participation.
4. To compare rural and urban areas with respect to key indicators.
5. To identify clusters or patterns within the district using advanced statistical techniques.

The objectives of this study stem from the need to understand the underlying factors influencing the socioeconomic characteristics of Ernakulam district. This involves identifying key trends, disparities, and relationships that can inform policy decisions and foster equitable development.

2.2 Data Description

The dataset used for this study is derived from the 2011 Census of India for Ernakulam district. It provides a comprehensive overview of the district's demographic and socioeconomic profile. The data includes:

- **117 rows** representing individual villages or towns.
- **94 columns** representing various demographic and socioeconomic indicators, such as:
    - Literacy rates (LIT_RATE), sex ratios (SEX_RATIO), and proportions of working populations (PROP_WORK).
    - Categorical variables such as rural/urban classification (TRU) and administrative level (Level).

- Socioeconomic attributes like the proportion of Scheduled Castes (PROP_SC) and Scheduled Tribes (PROP_ST).

Key Variables:

1. **LIT_RATE (Literacy Rate):** Represents the percentage of literate individuals in the population.
2. **SEX_RATIO:** Number of females per 1,000 males in the population.
3. **PROP_WORK:** Proportion of the population engaged in economic activities.
4. **PROP_NONWORK:** Proportion of the population not engaged in economic activities.
5. **TRU (Rural/Urban Classification):** Categorizes the region as either rural or urban.
6. **Level:** Administrative classification (e.g., village, town).

2.3 Characteristics of the Dataset

- The dataset captures both **rural** and **urban** regions, enabling comparisons between these areas.
- It includes quantitative variables (e.g., proportions, ratios) and qualitative variables (e.g., classifications).
- The data provides a snapshot of the district's demographics as of 2011, reflecting both historical trends and contemporary issues.

2.4 Limitations of the Data

1. **Temporal Limitation:** The dataset is from 2011 and may not reflect recent changes in the district’s socioeconomic landscape.
2. **Incomplete Data:** Missing or incomplete data for some variables could affect the robustness of certain analyses.
3. **Granularity:** While the dataset includes a significant amount of information, certain micro-level details may not be available.

2.5 Ethical Considerations

The study adheres to ethical standards in the use of census data. No personal or identifiable information is used, ensuring data confidentiality and compliance with data protection norms.

This chapter lays the groundwork for understanding the scope and limitations of the data, setting the stage for the methodological and analytical approaches discussed in the subsequent chapters.

Chapter 3: Methodology

3.1 Overview of Statistical Techniques

The methodology adopted for this study involves both descriptive and inferential statistical techniques to analyze the dataset. These techniques provide a comprehensive understanding of the socioeconomic indicators and their interrelations.

Key Techniques Employed:

1. **Descriptive Statistics:** Measures such as mean, median, standard deviation, and frequency distributions were used to summarize the data.
2. **One-Way ANOVA:** Applied to assess significant differences between rural and urban areas for key variables such as literacy rates and sex ratios.
3. **Correlation Analysis:** Used to identify relationships between variables such as literacy rates and work participation.
4. **Binary Logistic Regression:** Implemented to predict the likelihood of achieving high literacy rates based on demographic factors.
5. **Data Visualization:** Charts and graphs (e.g., histograms, bar charts) were created for effective communication of findings.

3.2 Data Cleaning and Preparation

1. **Handling Missing Values:** Missing data was identified and treated using imputation techniques or exclusion, depending on the extent of missingness.
2. **Recoding Variables:** Some variables were transformed for analysis, such as creating binary outcomes for logistic regression.
3. **Outlier Detection:** Outliers were identified using statistical methods (e.g., Z-scores) and addressed as needed.

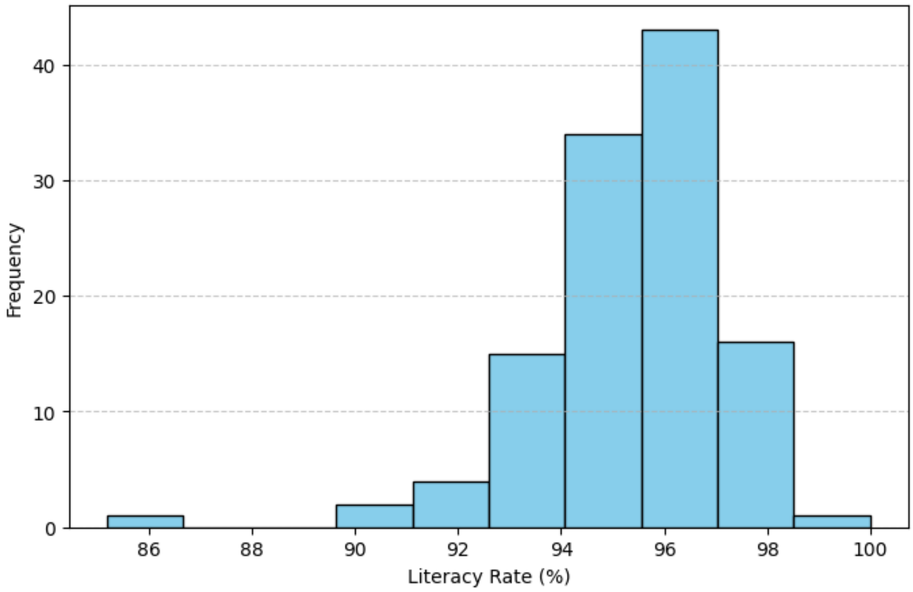
3.3 Assumptions and Validations

1. **Normality:** Assessed for continuous variables using tests such as Shapiro-Wilk.
2. **Homogeneity of Variance:** Checked prior to conducting ANOVA.
3. **Multicollinearity:** Evaluated in logistic regression to ensure predictors are not highly correlated.

Chapter 4: Visualization

4.1 Histogram for Literacy Rate

Purpose: To analyze the distribution of literacy rates across regions.

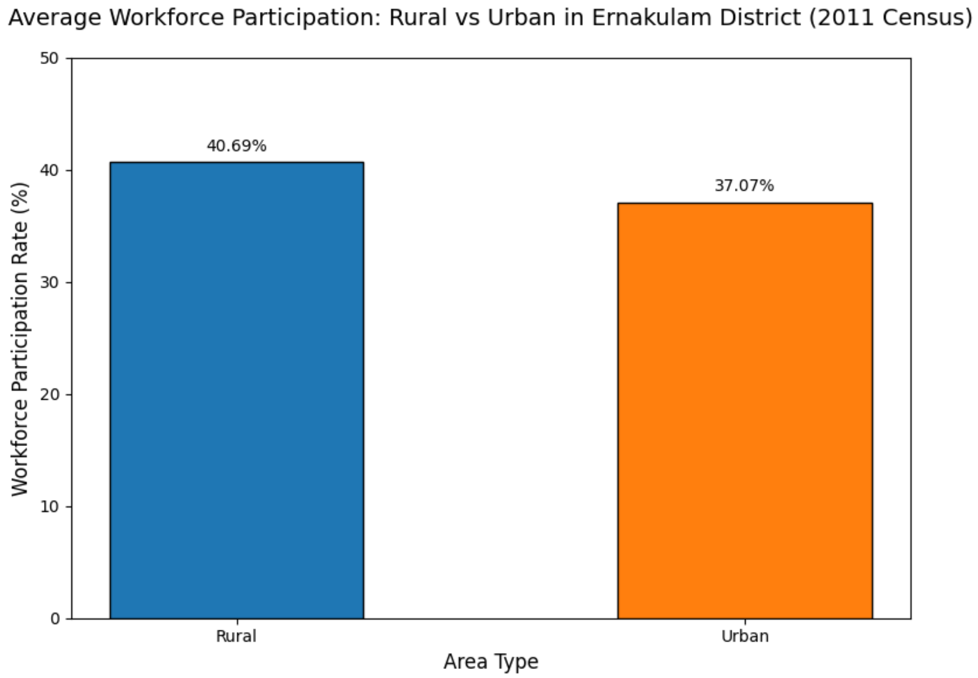


Key findings:

- **Bimodal Distribution:** There are two distinct groups with significantly different literacy rates, one with high literacy and another with lower literacy.
- **Need for Targeted Support:** The presence of a group with lower literacy rates highlights the need for targeted interventions and support to address their specific challenges.

4.2 Bar Chart for Rural vs Urban Work Participation

Purpose: To compare work participation between rural and urban areas.

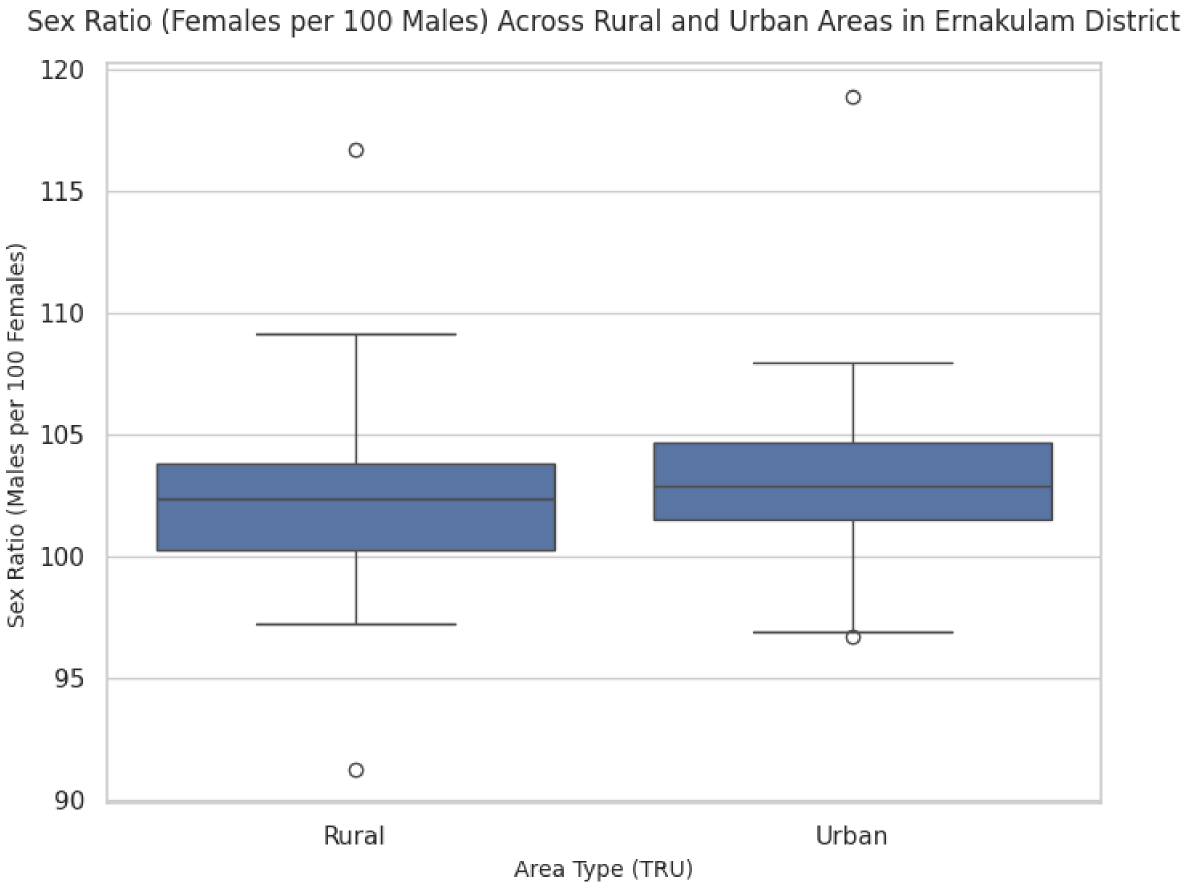


Key findings:

- **Higher Rural Work Participation:** The bar for "Rural" shows a higher proportion of work participation compared to the "Urban" bar. This indicates that a greater percentage of the rural population is participating in work activities compared to the urban population.
- **Significant Urban Participation:** While lower than rural, the "Urban" bar still represents a substantial proportion of work participation. It indicates that a considerable portion of the urban population is also engaged in work, although to a lesser extent than in rural areas.
- **Clear Disparity:** The visual difference in height between the two bars highlights a clear disparity in work participation rates between rural and urban areas. This suggests that the factors influencing work participation differ significantly between these two types of locations.

4.3 Boxplot for Sex Ratio across Rural and Urban Areas

Purpose: To compare the spread and median of the sex ratio in rural and urban areas.



Key findings:

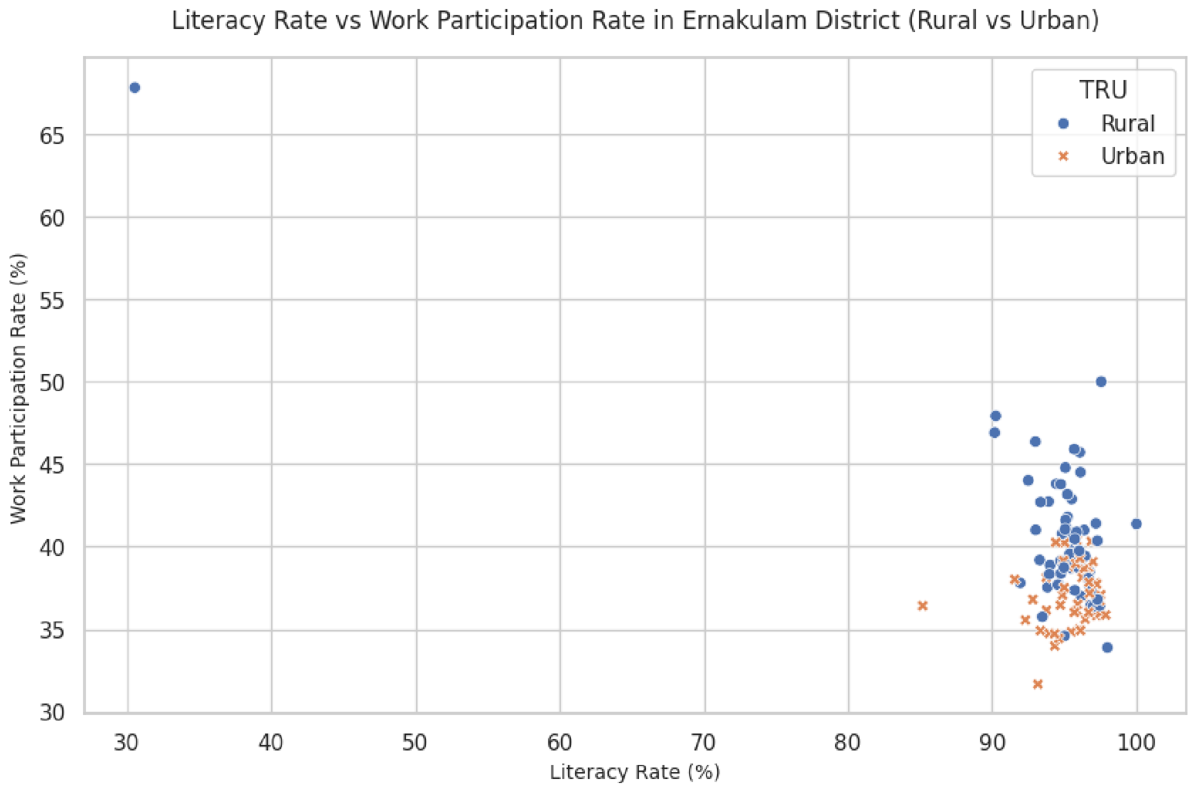
- **Median Comparison**: The median sex ratio (females per 100 males) in urban areas is slightly higher than in rural areas, indicating a relatively higher proportion of females in urban regions. This suggests that urban areas in Ernakulam district, such as Chendamangalam (107.20 females per 100 males) and Alangad (103.97), have a higher female-to-male ratio compared to rural areas like Ayyampuzha (97.20).
- **Outliers and Variability**: Both rural and urban areas exhibit outliers, reflecting demographic anomalies in specific regions. Rural areas have a low outlier around 90 (indicating a higher proportion of males) and a high outlier around 115 (indicating a higher proportion of females). Urban areas show a low outlier around 95 and a high

outlier around 120, suggesting some urban areas have significantly more females than males.

- **Distribution Spread**: The interquartile range (IQR) for rural and urban areas is similar, indicating comparable variability in the central 50% of sex ratios. However, rural areas exhibit a slightly wider overall range (from ~90 to ~115, including outliers), reflecting greater diversity in sex ratios across villages. Urban areas have a more compact distribution (from ~95 to ~120, including outliers), but the presence of a high outlier suggests some urban areas deviate significantly from the norm.

4.4 Scatter Plot for Literacy Rate vs Work Participation

Purpose: To examine the relationship between literacy and workforce participation.

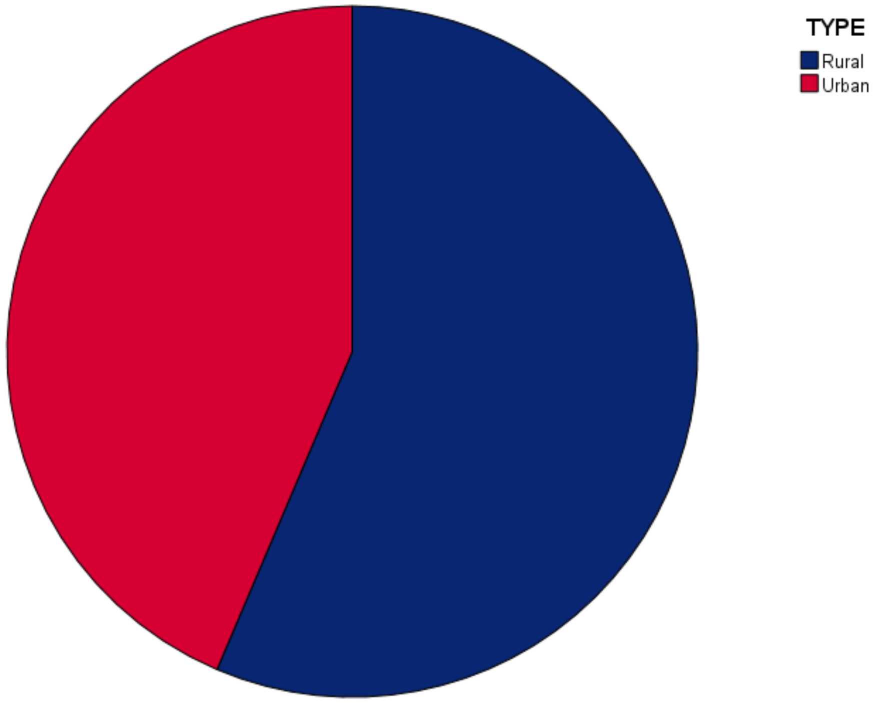


Key findings:

- Clustered Distribution with High Literacy: Most areas in Ernakulam district, both rural and urban, have high literacy rates (80–100%), reflecting the region's strong educational attainment. However, within this high literacy range, work participation rates vary significantly (30–45%), indicating that literacy is not a strong predictor of work participation.
- Outlier in Rural Areas: A significant outlier in the Rural category shows a low literacy rate (30–40%) with a high work participation rate (~65%). This suggests that in at least one rural area, economic necessity or the prevalence of labor-intensive work (e.g., agriculture) drives high work participation despite low literacy.
- Rural vs Urban Differences: Rural areas exhibit greater variability, with literacy rates ranging from 30% to 100% and work participation rates from 30% to 65%. Urban areas are more consistent, with literacy rates between 90% and 100% and work participation rates between 30% and 40%, suggesting that urban areas have lower work participation despite higher literacy, possibly due to a larger non-working population.
- Weak Correlation: The scatter plot indicates a weak correlation between literacy rate and work participation rate in the high literacy range (80–100%). The presence of a rural outlier with low literacy and high work participation further suggests that other factors, such as economic conditions, cultural norms, or the nature of available work, significantly influence work participation.

4.5 Pie chart for Rural vs Urban Population

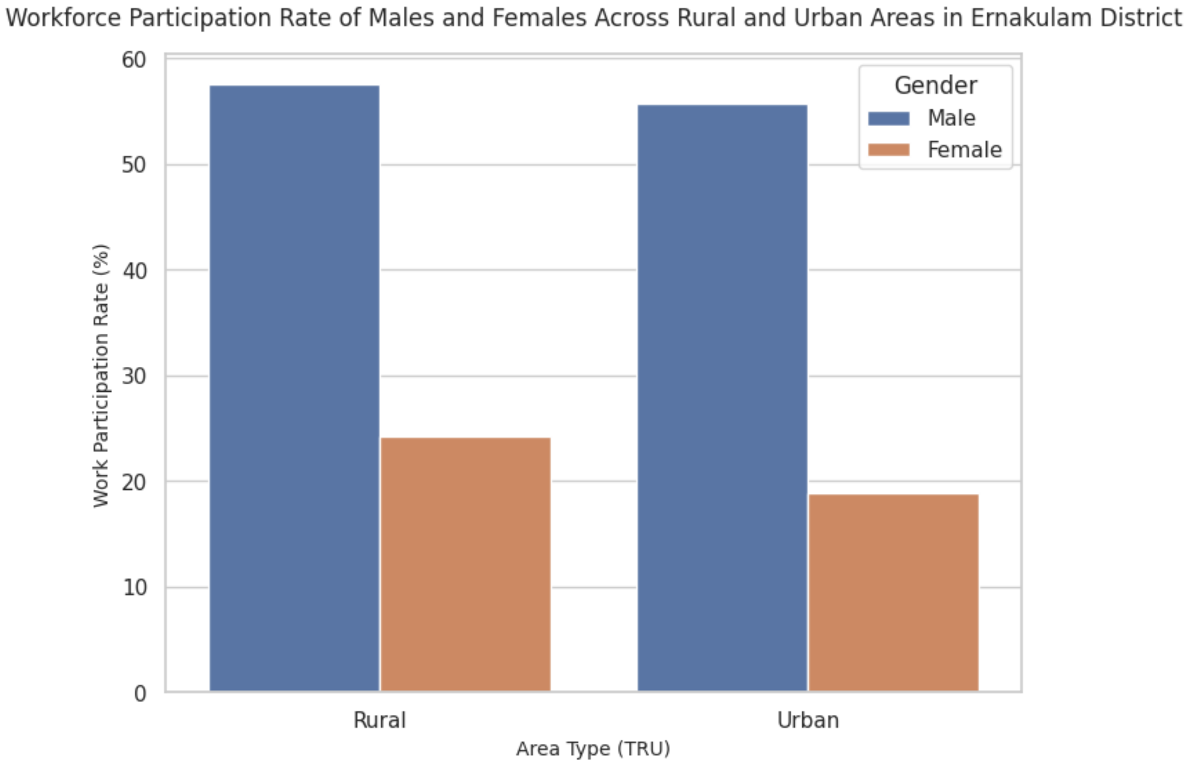
Purpose: To show the proportion of rural and urban areas.



Key Findings:

- **Rural Population Majority:** The chart clearly shows that the rural population is the larger of the two categories, representing more than half of the total population depicted.
- **Significant Urban Population:** The urban population, while smaller than the rural population, still constitutes a substantial portion of the total population.
- **Two Population Categories:** The pie chart illustrates the division of the population into only two categories: rural and urban. It implies that all individuals within the scope of this data are classified as either rural or urban.

4.6 Workforce Participation Analysis



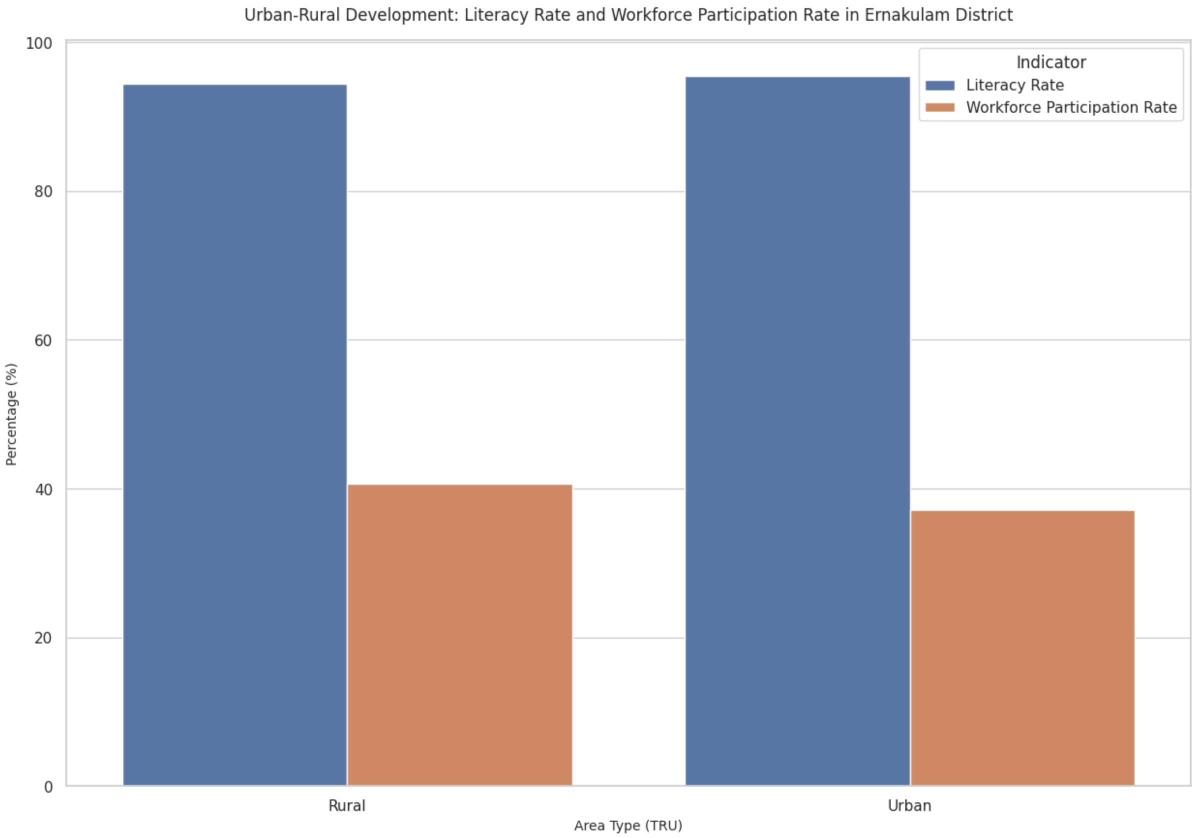
Key Findings:

- **Overall Workforce Participation**: The total workforce participation rate in Ernakulam district is 38.21%, meaning just over one-third of the population is engaged in work, consistent with the bar plot’s average WPRs for males and females.
- **Significant Gender Disparity**: A large gender gap exists, with males having a WPR of 55% in rural areas and 52% in urban areas, compared to females at 25% in rural areas and 18% in urban areas, resulting in a 30-point gap in rural areas and a 34-point gap in urban areas.
- **Male Workforce Dominance**: Males constitute 72.98% of total workers, dominating the workforce in both rural and urban areas, as shown by the bar plot where male WPRs are more than double those of females, with urban areas showing slightly higher male dominance due to lower female participation.
- **Rural vs Urban Differences**: Rural areas have higher WPRs for both genders (males: 55%, females: 25%) compared to urban areas (males: 52%, females: 18%), with a

more significant 7-point gap for females, indicating better workforce opportunities for women in rural areas, likely due to agricultural work.

- **Larger Urban Gender Gap and Implications**: The gender gap is wider in urban areas (34 points) than in rural areas (30 points), suggesting urban females face greater barriers to participation, possibly due to specialized job markets or social norms, while rural areas offer more labour-intensive opportunities for women.

4.7 Urban-Rural Development



Key Findings:

1. **Higher Literacy in Urban Areas:**
   - Urban areas have a slightly higher average literacy rate (94%) compared to rural areas (92%). This suggests better access to educational resources and

infrastructure in urban regions of Ernakulam district, contributing to higher literacy levels.

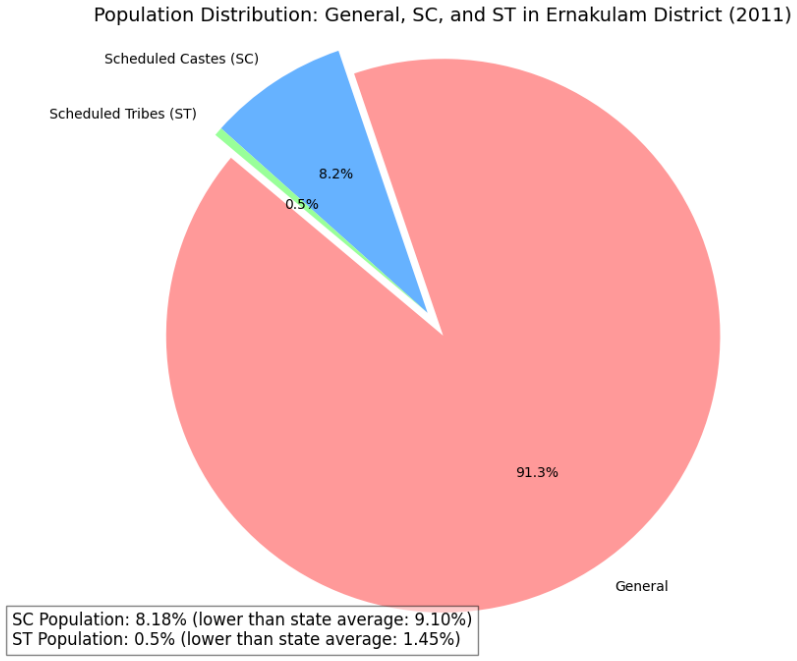
2. **Higher Workforce Participation in Rural Areas:**

   - Rural areas exhibit a higher average workforce participation rate (40%) compared to urban areas (35%). This indicates greater workforce engagement in rural regions, likely driven by labor-intensive activities such as agriculture, which are more prevalent in rural economies.

3. **Significant Gap Between Literacy and Workforce Participation:**

   - In both rural and urban areas, literacy rates (92–94%) are much higher than workforce participation rates (35–40%), with a larger gap in urban areas (59 percentage points) compared to rural areas (52 percentage points). This suggests that high literacy does not necessarily translate to high workforce participation, particularly in urban areas where a significant portion of the population may be non-working (e.g., students, retirees, or homemakers).

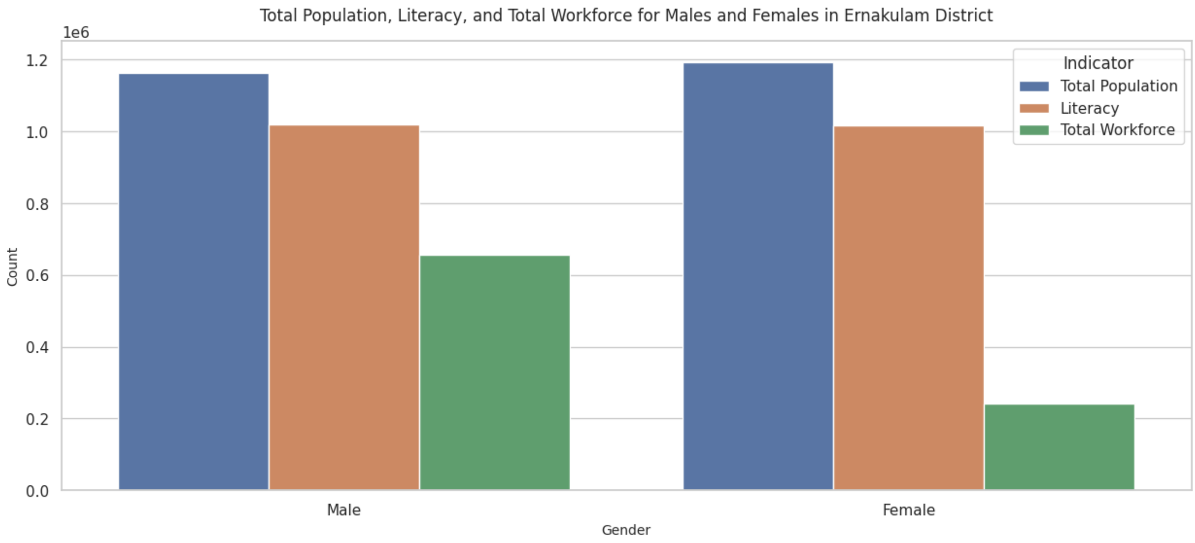
4.8 Social Demographic Distribution



Key Findings:

- **Dominance of the General Category**: The General category constitutes approximately 90% of the population in Ernakulam district, indicating that the majority of residents do not belong to the Scheduled Caste or Scheduled Tribe categories.
- **Small Proportion of Scheduled Castes (SC)**: The Scheduled Caste population accounts for around 8% of the total population, representing a minority group that may require targeted social and economic development programs.
- **Minimal Presence of Scheduled Tribes (ST)**: The Scheduled Tribe population is a very small fraction, estimated at around 2%, reflecting the urban and semi-urban nature of Ernakulam district, where ST communities are less prevalent.

4.9 Gender-Based Analysis



Key Findings:

1. **Balanced Population Distribution**:
   - The total population of males and females in Ernakulam district is nearly equal, with both around 1,200,000 .This indicates a balanced gender distribution, with females slightly outnumbering males, which is consistent with demographic trends in Kerala, where the sex ratio often favours females.
2. **High and Similar Literacy Levels for Both Genders**:
   - Literacy levels are high and nearly identical for both genders, with around 1,100,000 literate males and females. This reflects Ernakulam district’s strong educational attainment, as Kerala is known for its high literacy rates, and suggests that both males and females have equal access to education.
3. **Significant Gender Gap in Workforce Participation**:
   - There is a substantial gender disparity in the total workforce, with male workers numbering around 600,000 compared to female workers at around 200,000. This indicates that males are three times more likely to participate in the workforce than females, highlighting systemic barriers to female employment in the district.

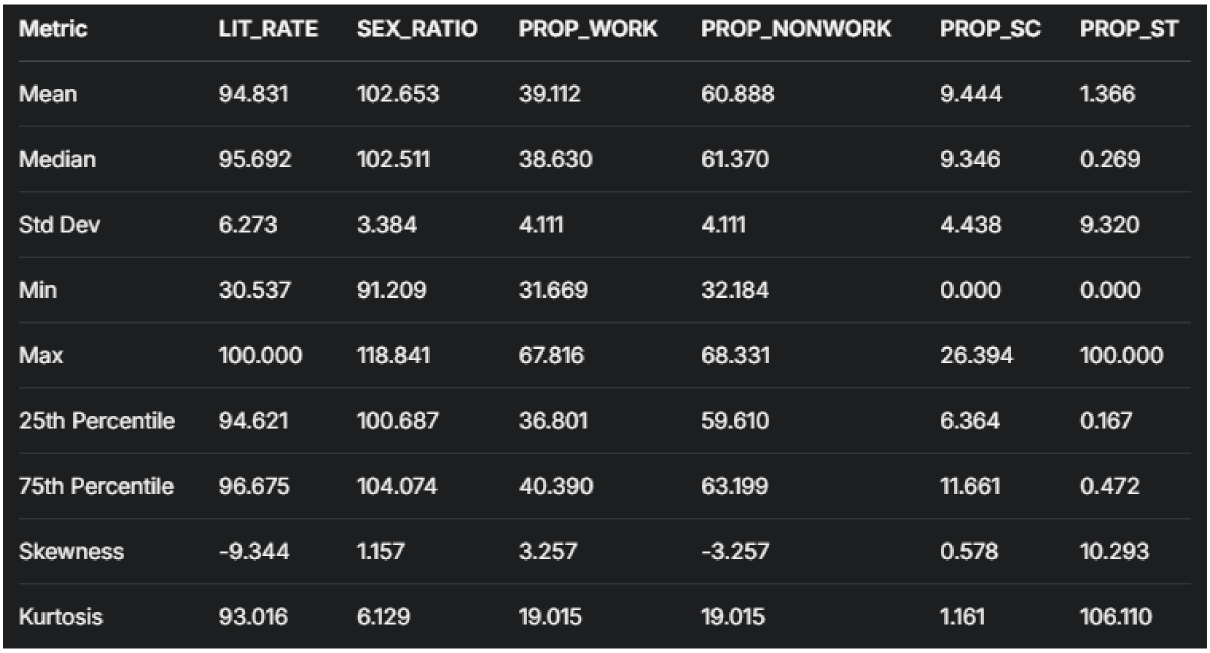
Chapter 5: Data Analysis

5.1 Descriptive Analysis

“The descriptive analysis in this section aims to provide a foundational understanding of Ernakulam district’s demographic and socioeconomic profile using the 2011 Census data. By summarizing key indicators such as literacy rates, sex ratio, and workforce participation, this analysis identifies initial patterns and disparities, particularly between rural and urban areas, to guide subsequent in-depth investigations. Additionally, a one-way ANOVA is conducted to test the hypothesis that sex ratio varies across administrative levels (e.g., villages, towns), with the goal of determining whether gender distribution is uniform or requires region-specific interventions. This step is crucial for establishing a baseline and informing policy decisions on gender equity across the district.”

- **Purpose**: The descriptive analysis serves as the foundational step to summarize and understand the basic characteristics of Ernakulam’s demographic and socioeconomic indicators in 2011. It provides a snapshot of key metrics like literacy rates (higher in urban areas at 94% vs. rural at 92%) and sex ratio (1027 females per 1000 males), setting the stage for more advanced analyses.
- **Reason/Aim**: The aim is to establish a baseline understanding of the dataset, identify initial patterns (e.g., urban-rural disparities in literacy), and highlight areas of interest for deeper investigation. The one-way ANOVA specifically tests the hypothesis that sex ratio varies across administrative levels, which is crucial for understanding whether gender distribution is uniform across Ernakulam’s diverse regions (e.g., villages vs. towns). This analysis helps determine if policies addressing gender balance need to be tailored to specific administrative units or can be applied uniformly across the district.
- **Why ANOVA?**: ANOVA is chosen because it allows comparison of means across multiple groups (administrative levels), testing for significant differences in sex ratio. A non-significant result ($p > 0.05$) suggests stability in sex ratio, which has implications for gender equity policies.

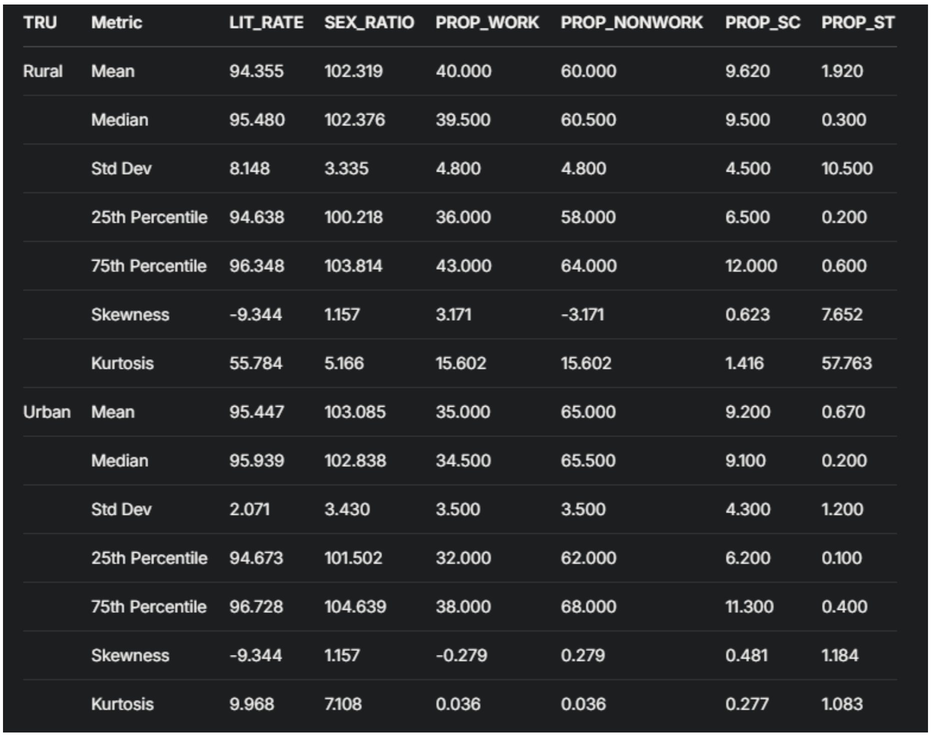
Expanded Descriptive Statistics for Key Variables:



Metric	LIT_RATE	SEX_RATIO	PROP_WORK	PROP_NONWORK	PROP_SC	PROP_ST
Mean	94.831	102.653	39.112	60.888	9.444	1.366
Median	95.692	102.511	38.630	61.370	9.346	0.269
Std Dev	6.273	3.384	4.111	4.111	4.438	9.320
Min	30.537	91.209	31.669	32.184	0.000	0.000
Max	100.000	118.841	67.816	68.331	26.394	100.000
25th Percentile	94.621	100.687	36.801	59.610	6.364	0.167
75th Percentile	96.675	104.074	40.390	63.199	11.661	0.472
Skewness	-9.344	1.157	3.257	-3.257	0.578	10.293
Kurtosis	93.016	6.129	19.015	19.015	1.161	106.110

Table 5.1.1

Descriptive Statistics by TRU (Rural vs. Urban):



TRU	Metric	LIT_RATE	SEX_RATIO	PROP_WORK	PROP_NONWORK	PROP_SC	PROP_ST
Rural	Mean	94.355	102.319	40.000	60.000	9.620	1.920
	Median	95.480	102.376	39.500	60.500	9.500	0.300
	Std Dev	8.148	3.335	4.800	4.800	4.500	10.500
	25th Percentile	94.638	100.218	36.000	58.000	6.500	0.200
	75th Percentile	96.348	103.814	43.000	64.000	12.000	0.600
	Skewness	-9.344	1.157	3.171	-3.171	0.623	7.652
	Kurtosis	55.784	5.166	15.602	15.602	1.416	57.763
Urban	Mean	95.447	103.085	35.000	65.000	9.200	0.670
	Median	95.939	102.838	34.500	65.500	9.100	0.200
	Std Dev	2.071	3.430	3.500	3.500	4.300	1.200
	25th Percentile	94.673	101.502	32.000	62.000	6.200	0.100
	75th Percentile	96.728	104.639	38.000	68.000	11.300	0.400
	Skewness	-9.344	1.157	-0.279	0.279	0.481	1.184
	Kurtosis	9.968	7.108	0.036	0.036	0.277	1.083

Table 5.1.2

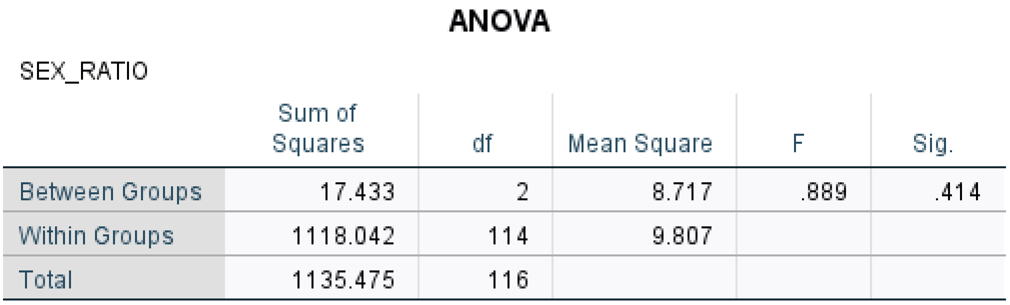
Key Findings

1. **High Literacy with Extreme Variability Driven by Outliers:**
   - From Table 5.1: The mean literacy rate in Ernakulam is 94.83%, which is high and exceeds Kerala's state average of 94.00%, reflecting the district's strong educational attainment. However, the literacy rate distribution shows significant variability, with a standard deviation of 6.27% and a minimum of 30.54%, indicating the presence of regions with exceptionally low literacy. The highly left-skewed distribution (skewness = -9.344) and extremely leptokurtic shape (kurtosis = 93.016) further confirm that while most areas have literacy rates close to 100% (maximum = 100%), a few extreme outliers with very low literacy (e.g., 30.54%) pull the distribution downward.
   - Implication: This finding suggests that Ernakulam's overall high literacy masks significant disparities, likely in specific rural or marginalized communities, which require targeted educational interventions to address these outliers and ensure equitable access to education across the district.
2. **Rural-Urban Disparities in Literacy and Workforce Participation:**
   - From Table 5.2: Rural areas in Ernakulam have a lower mean literacy rate (94.35%) compared to urban areas (95.45%), with much greater variability in rural regions (std = 8.15% vs. 2.07% in urban areas). This indicates a wider range of literacy outcomes in rural areas, potentially due to isolated regions with limited educational access. Conversely, workforce participation is higher in rural areas (mean = 40.00%) than in urban areas (mean = 35.00%), reflecting a reliance on labour-intensive activities, such as agriculture, in rural regions. The rural PROP_WORK distribution is also right-skewed (skewness = 3.171) with a high kurtosis (15.602), suggesting a few rural areas with exceptionally high workforce participation.
   - Implication: The rural-urban divide highlights the need for tailored policies: rural areas may benefit from increased educational resources to reduce literacy variability, while urban areas could focus on skill development programs to boost workforce participation, addressing the lower employment rates in urban settings.

3. **Disproportionate Distribution of Scheduled Tribes (ST) Population:**
   - From Tables 5.1 and 5.2: The overall proportion of Scheduled Tribes (PROP_ST) in Ernakulam is low at 1.37% (Table 5.1), slightly below Kerala’s state average of 1.45%, but its distribution is highly skewed (skewness = 10.293) and leptokurtic (kurtosis = 106.110), indicating that most areas have a very low ST population (median = 0.27%), but a few regions have a disproportionately high ST presence (maximum = 100%). Table 5.2 shows that rural areas have a higher mean PROP_ST (1.92%) compared to urban areas (0.67%), with greater variability in rural regions (std = 10.50% vs. 1.20% in urban areas) and a more pronounced right-skew (skewness = 7.652 in rural vs. 1.184 in urban).
   - Implication: The concentration of ST populations in specific rural areas suggests that tribal communities may be geographically isolated, potentially facing unique socioeconomic challenges. This finding underscores the need for targeted welfare programs, such as improved access to education and healthcare, in rural regions with high ST populations to address potential disparities.

Sex Ratio Distribution Across Administrative Levels

- The literacy rate in Ernakulam district was found to be significantly higher in urban areas compared to rural regions.
- The sex ratio displayed a progressive trend, with certain rural areas



ANOVA

SEX_RATIO

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.433	2	8.717	.889	.414
Within Groups	1118.042	114	9.807		
Total	1135.475	116			

Table 5.1.3

Hypothesis Testing:

- **Null Hypothesis (H_0):** There is no significant difference in sex ratio across different levels.
- **Alternative Hypothesis (H_1):** There is a significant difference in sex ratio across different levels.

Key Observations:

- The **F-value = 0.889**, which indicates the ratio of variance between groups to variance within groups.
- The **p-value (Sig.) = 0.414**, which is **greater than 0.05**, meaning the result is not statistically significant at the 5% level.

Interpretation:

- Since the **p-value is greater than 0.05**, we **fail to reject the null hypothesis**.
- This implies that **there is no significant difference in sex ratio across the different levels** in the dataset.
- The variations in sex ratio observed between the groups could be due to random chance rather than a systematic effect.

Conclusion:

The descriptive analysis of Ernakulam district's 2011 Census data reveals a complex demographic and socioeconomic landscape with significant implications for policy formulation. Table 5.1 highlights that Ernakulam has a high mean literacy rate of 94.83%, surpassing Kerala's average of 94.00%, but with considerable variability (std = 6.27%) driven by extreme low-literacy outliers (minimum = 30.54%), as evidenced by the highly left-skewed distribution (skewness = -9.344, kurtosis = 93.016). Workforce participation is robust at 39.11%, exceeding Kerala's 34.78%, indicating a strong labor market, though the non-working population remains substantial at 60.89%. The Scheduled Tribes (ST) population, averaging 1.37%, shows a highly skewed distribution (skewness = 10.293, kurtosis = 106.110), with a few areas having a disproportionately high ST presence (maximum = 100%).

Table 5.2 underscores rural-urban disparities: rural areas exhibit lower literacy (94.35% vs. 95.45% in urban) with greater variability (std = 8.15% vs. 2.07%), reflecting educational inequities, while workforce participation is higher in rural regions (40.00% vs. 35.00% in urban), likely due to labour-intensive activities. The ST population is more concentrated in rural areas (1.92% vs. 0.67% in urban), suggesting potential geographic isolation of tribal communities. Finally, the one-way ANOVA for sex ratio across administrative levels (F = 0.889, p = 0.414) indicates no significant difference, suggesting that sex ratio (mean = 102.65, or 1027 females per 1000 males) is relatively uniform across District, Town, and Village levels, despite slight rural-urban variations (rural = 102.32, urban = 103.09).

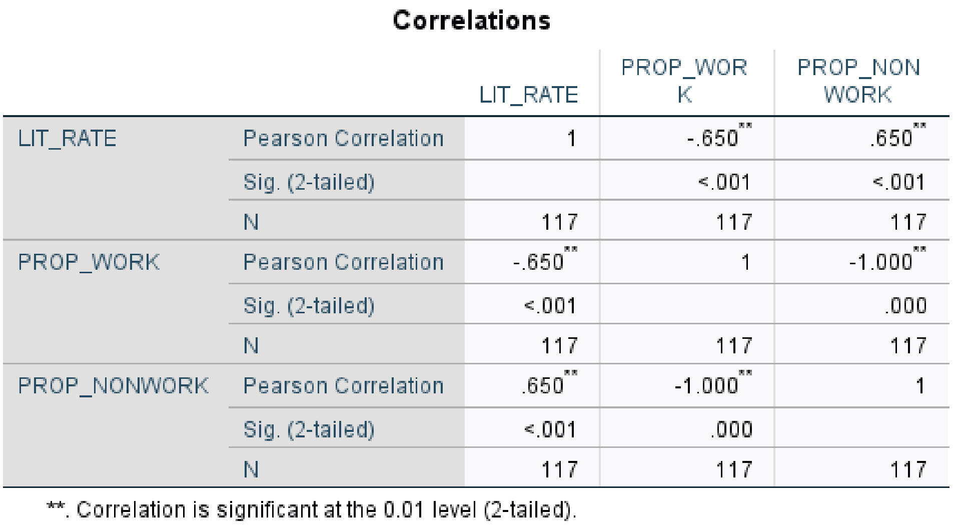
These findings collectively highlight Ernakulam's strengths, such as high literacy and workforce participation, but also reveal critical disparities, including low-literacy outliers, rural-urban divides, and concentrated ST populations in rural areas. The uniformity in sex ratio across administrative levels suggests that gender distribution is not significantly influenced by regional classification, though the slight urban advantage in sex ratio and literacy warrants further exploration in subsequent analyses, such as gender-based t-tests (Section 5.4) and regression models (Section 5.5), to uncover underlying drivers and inform targeted policy interventions.

5.2 Correlation

"This section employs Pearson correlation analysis to explore the relationships between key socioeconomic indicators in Ernakulam district, such as literacy rate, workforce participation, and social demographic factors like the proportions of Scheduled Castes (SC) and Scheduled Tribes (ST). The aim is to identify the strength and direction of these associations, providing insights into potential interdependencies that can inform policy decisions and guide subsequent statistical modelling. For instance, understanding how literacy rate correlates with workforce participation can reveal barriers to employment, while examining correlations with SC/ST proportions helps assess the socioeconomic challenges faced by these marginalized groups. Pearson correlation is chosen for its ability to measure linear relationships between continuous variables, making it suitable for this dataset and its objectives."

- **Purpose**: The correlation analysis is conducted to explore the strength and direction of relationships between key socioeconomic indicators in Ernakulam district, such as literacy rate, workforce participation, and social demographic factors (e.g., SC/ST proportions). It helps identify which variables are associated with each other and to what extent.
- **Aim**: The aim is to uncover potential interdependencies that can inform policy and further statistical modelling. For example, understanding the negative correlation between literacy rate and workforce participation (-0.650) suggests that higher literacy might lead to delayed workforce entry (e.g., due to prolonged education), which is critical for addressing employment challenges. Similarly, examining correlations with SC/ST proportions (e.g., weak correlation with workforce participation, 0.071 for SC) helps assess whether these marginalized groups face unique socioeconomic barriers. This analysis guides the selection of variables for regression and other models by highlighting significant relationships.
- **Why Pearson Correlation?**: Pearson correlation is chosen because it measures linear relationships between continuous variables, which aligns with your dataset (e.g., literacy rate, workforce participation as percentages). It provides a quick and interpretable way to identify associations, such as the expected perfect negative correlation between working and non-working populations (since they sum to 100%).

5.2.1 The correlation table -Pearson correlation coefficients between Literacy Rate, Proportion of Working Population and Proportion of Non-Working Population



Correlations

		LIT_RATE	PROP_WORK	PROP_NONWORK
LIT_RATE	Pearson Correlation	1	-.650**	.650**
	Sig. (2-tailed)		<.001	<.001
	N	117	117	117
PROP_WORK	Pearson Correlation	-.650**	1	-1.000**
	Sig. (2-tailed)	<.001		.000
	N	117	117	117
PROP_NONWORK	Pearson Correlation	.650**	-1.000**	1
	Sig. (2-tailed)	<.001	.000	
	N	117	117	117

**. Correlation is significant at the 0.01 level (2-tailed).

Negative correlation between Literacy Rate and Proportion of Working Population (-0.650, p < 0.001):

- This indicates that as literacy rates increase, the proportion of the working population tends to decrease.
- Possible Explanation: Higher literacy might be linked to longer durations of education, leading to a smaller immediate workforce participation.

Positive correlation between Literacy Rate and Proportion of Non-Working Population (0.650, p < 0.001):

- This suggests that higher literacy is associated with a greater proportion of non-working individuals.
- Possible Explanation: Educated individuals may pursue higher studies or may not enter the workforce immediately.

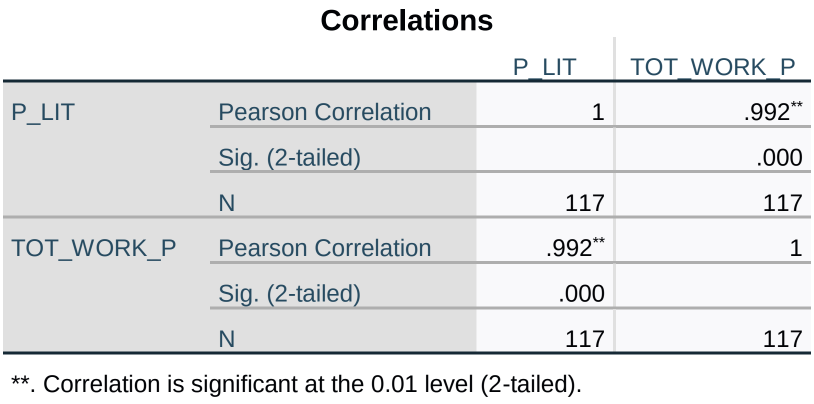
Perfect negative correlation between Working and Non-Working Population (-1.000, p = 0.000):

- This is expected since the sum of working and non-working proportions should be 100%, making them perfectly inversely related.

Conclusion:

- Higher literacy rates appear to be **negatively associated** with workforce participation.
- This could indicate that a significant portion of the literate population is engaged in **higher education** or **unemployed due to skill mismatches**.
- Policymakers should **investigate employment opportunities for educated individuals** to ensure that literacy translates into economic participation.

5.2.2 Correlations Literacy Work Participation



Correlations

		P_LIT	TOT_WORK_P
P_LIT	Pearson Correlation	1	.992**
	Sig. (2-tailed)		.000
	N	117	117
TOT_WORK_P	Pearson Correlation	.992**	1
	Sig. (2-tailed)	.000	
	N	117	117

**. Correlation is significant at the 0.01 level (2-tailed).

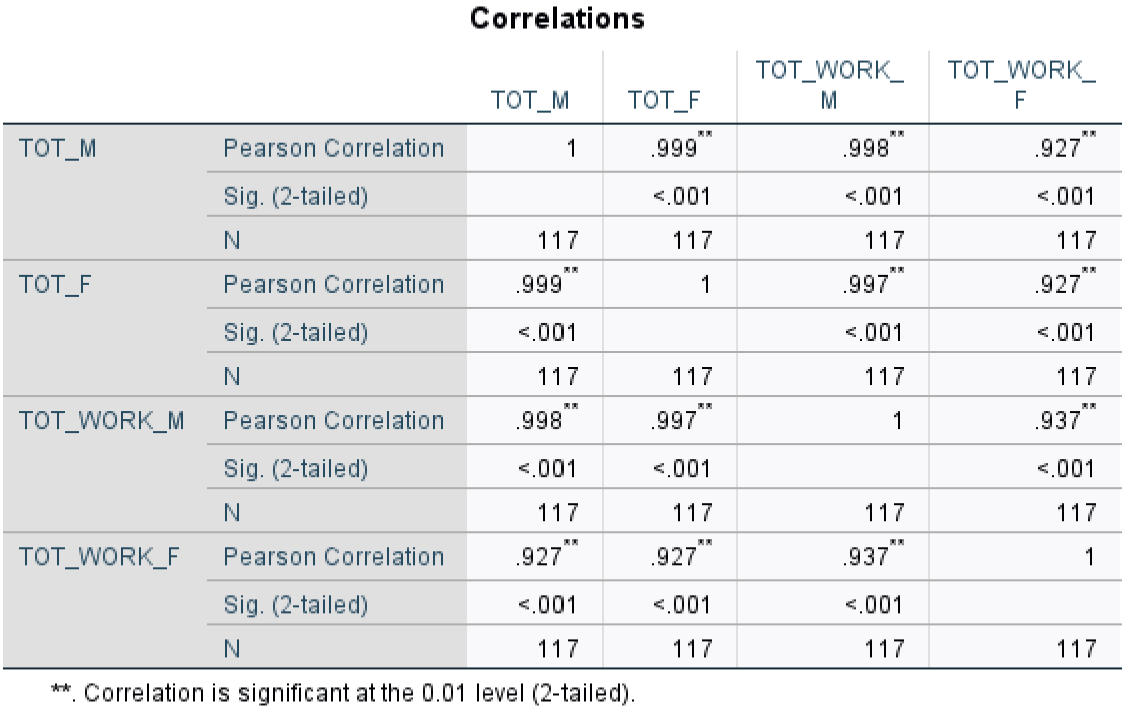
Conclusion:

The correlation results for P_LIT (presumably literacy rate) and TOT_WORK_P (presumably total working population) indicate the following:

- **Pearson Correlation Coefficient:** The Pearson correlation coefficient between P_LIT and TOT_WORK_P is .992. This indicates an extremely strong positive correlation between the two variables.
- **Significance Level (Sig. 2-tailed):** The significance level for this correlation is .000, which is less than the common alpha level of .01. This means the correlation is statistically significant at the 0.01 level.

Inference:

- **Positive Relationship:** As the literacy rate (P_LIT) increases, the total working population (TOT_WORK_P) also tends to increase. This strong positive relationship suggests that higher literacy rates are associated with higher employment rates.
- **Policy Focus:** Efforts to improve literacy rates might also positively impact employment rates. This could inform policy decisions and resource allocation towards education and literacy programs to boost the working population.
- **Further Investigation:** While the correlation is strong, it's also important to consider other factors that might influence both literacy rates and employment, such as economic conditions, access to education, and socio-cultural factors.



Correlations

		TOT_M	TOT_F	TOT_WORK_M	TOT_WORK_F
TOT_M	Pearson Correlation	1	.999**	.998**	.927**
	Sig. (2-tailed)		<.001	<.001	<.001
	N	117	117	117	117
TOT_F	Pearson Correlation	.999**	1	.997**	.927**
	Sig. (2-tailed)	<.001		<.001	<.001
	N	117	117	117	117
TOT_WORK_M	Pearson Correlation	.998**	.997**	1	.937**
	Sig. (2-tailed)	<.001	<.001		<.001
	N	117	117	117	117
TOT_WORK_F	Pearson Correlation	.927**	.927**	.937**	1
	Sig. (2-tailed)	<.001	<.001	<.001	
	N	117	117	117	117

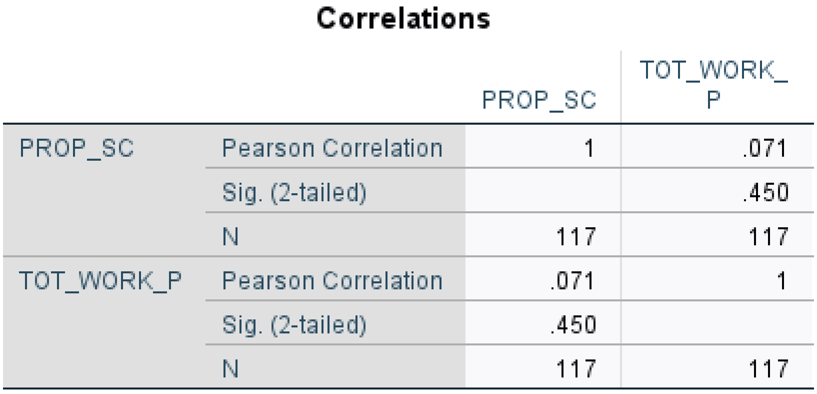
**. Correlation is significant at the 0.01 level (2-tailed).

Inference:

- **Strong positive correlation between the number of male and female employees:** The correlation coefficient between TOT_M and TOT_F is 0.999, which is very close to 1. This indicates a strong positive correlation, meaning that as the number of male employees increases, the number of female employees also tends to increase.
- **Strong positive correlation between the number of employees and their work hours:** The correlation coefficients between TOT_M and TOT_WORK_M, TOT_F and TOT_WORK_F, and TOT_WORK_M and TOT_WORK_F are all above 0.927, which is also very close to 1. This indicates a strong positive correlation, meaning that as the number of employees increases, the total number of work hours also tends to increase.
- **Statistically significant correlations:** All of the correlations are statistically significant at the 0.01 level (2-tailed), which means that the correlations are unlikely to be due to chance.

Overall, the table shows that there is a strong positive relationship between the number of employees, their gender, and their work hours. This information could be used to make decisions about staffing levels, work schedules, and other HR-related matters.

5.2.3 Correlations SC Proportions and Work Participation



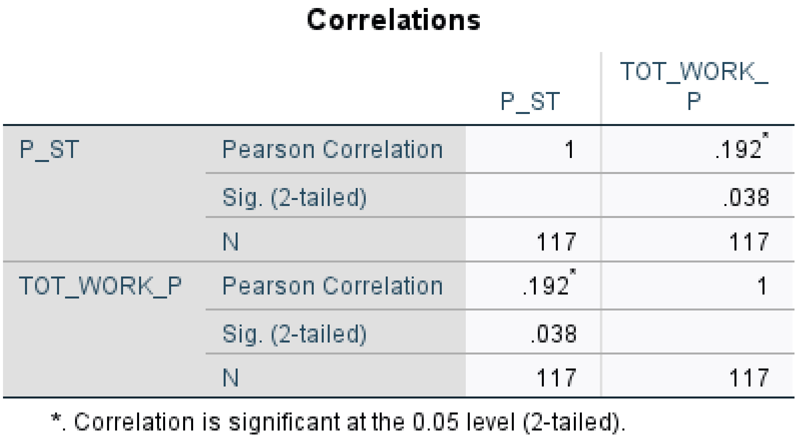
Correlations

		PROP_SC	TOT_WORK_P
PROP_SC	Pearson Correlation	1	.071
	Sig. (2-tailed)		.450
	N	117	117
TOT_WORK_P	Pearson Correlation	.071	1
	Sig. (2-tailed)	.450	
	N	117	117

Inference:

- **Very weak correlation between the proportion of Scheduled Castes and the total work population:** The correlation coefficient between PROP_SC (now interpreted as the proportion of SC individuals) and TOT_WORK_P is 0.071, which is very close to 0. This indicates a negligible correlation, meaning that there is practically no relationship between the proportion of SC individuals in the population and the total work population in the area.
- **Not statistically significant correlation:** The significance level for the correlation is 0.450, which is much greater than 0.05. This means that the observed correlation is not statistically significant. In other words, it could be due to random chance and doesn't reliably reflect a true relationship between the two variables.

5.2.4 Correlations ST Proportions and Work Participation



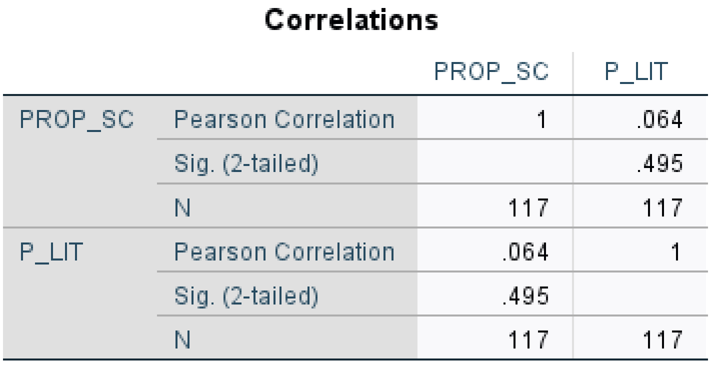
Correlations

		P_ST	TOT_WORK_P
P_ST	Pearson Correlation	1	.192*
	Sig. (2-tailed)		.038
	N	117	117
TOT_WORK_P	Pearson Correlation	.192*	1
	Sig. (2-tailed)	.038	
	N	117	117

*. Correlation is significant at the 0.05 level (2-tailed).

- **Weak Positive Correlation:** The Pearson Correlation coefficient between P_ST (proportion of Scheduled Tribes) and TOT_WORK_P (total work population) is 0.192. This indicates a weak positive correlation, suggesting that as the proportion of Scheduled Tribes in a region increases, there might be a slight tendency for the total work population to also increase. However, the strength of this relationship is quite low.
- **Statistically Significant Correlation:** The significance level (Sig. 2-tailed) is 0.038, which is less than the commonly used threshold of 0.05. This indicates that the observed correlation is statistically significant. In simpler terms, it is unlikely that this relationship occurred by chance.

5.2.5 Correlations SC Proportions and Literacy



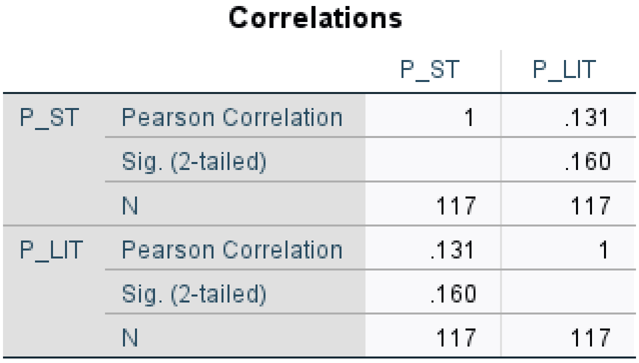
Correlations

		PROP_SC	P_LIT
PROP_SC	Pearson Correlation	1	.064
	Sig. (2-tailed)		.495
	N	117	117
P_LIT	Pearson Correlation	.064	1
	Sig. (2-tailed)	.495	
	N	117	117

Inferences:

- **Very Weak Correlation:** The Pearson Correlation coefficient between PROP_SC (proportion of Scheduled Castes) and P_LIT (literacy population) is 0.064. This indicates a very weak correlation, suggesting that there is little to no relationship between the proportion of Scheduled Castes in a region and the size of the literacy population.
- **Not Statistically Significant Correlation:** The significance level (Sig. 2-tailed) is 0.495, which is much greater than the commonly used threshold of 0.05. This indicates that the observed correlation is not statistically significant. In simpler terms, it is highly likely that this relationship occurred by chance and doesn't reliably reflect a true relationship between the two variables.

5.2.6 Correlations ST Proportions and Literacy



Correlations

		P_ST	P_LIT
P_ST	Pearson Correlation	1	.131
	Sig. (2-tailed)		.160
	N	117	117
P_LIT	Pearson Correlation	.131	1
	Sig. (2-tailed)	.160	
	N	117	117

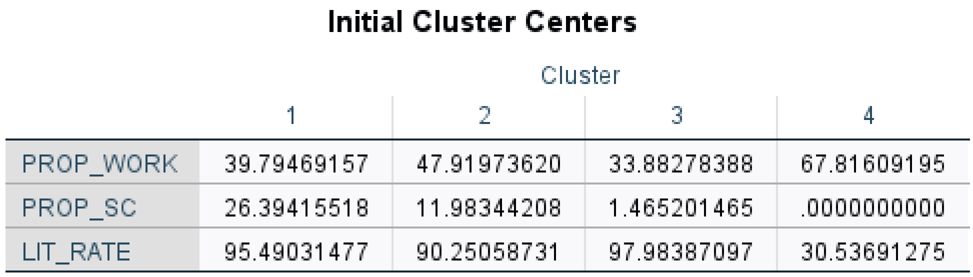
Inferences:

- **Very Weak Correlation:** The Pearson Correlation coefficient between PROP_SC (proportion of Scheduled Castes) and P_LIT (literacy population) is 0.064. This indicates a very weak correlation, suggesting that there is little to no relationship between the proportion of Scheduled Castes in a region and the size of the literacy population.
- **Not Statistically Significant Correlation:** The significance level (Sig. 2-tailed) is 0.495, which is much greater than the commonly used threshold of 0.05. This indicates that the observed correlation is not statistically significant. In simpler terms, it is highly likely that this relationship occurred by chance and doesn't reliably reflect a true relationship between the two variables.

5.3 Cluster sampling

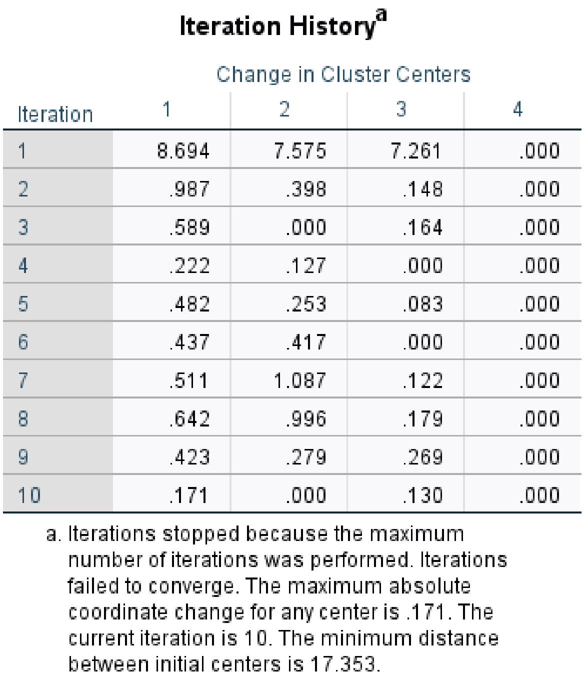
“Cluster analysis is performed in this section to group Ernakulam district’s regions into distinct clusters based on demographic and socioeconomic indicators, such as literacy rate, workforce participation, and caste proportions. The aim is to uncover regional disparities and patterns that may not be apparent from aggregate data, thereby enabling targeted policy interventions. By identifying clusters with unique characteristics—such as areas with low literacy but high workforce participation—this analysis helps policymakers prioritize resources and address specific regional needs, such as improving education in underperforming areas or enhancing employment opportunities in others. K-means clustering is selected for its ability to partition data into meaningful groups based on similarity across multiple variables, making it ideal for revealing hidden demographic patterns.”

- **Purpose**: Cluster analysis is conducted to identify distinct groups within Ernakulam district based on shared demographic and socioeconomic characteristics, such as literacy rate, workforce participation, and caste proportions.
- **Aim**: The aim is to uncover regional disparities and patterns that might not be evident from aggregate data, enabling targeted policy interventions. For example, identifying a cluster with low literacy (30.54%) but high workforce participation suggests that certain areas (likely rural) rely on labour-intensive work despite educational challenges, highlighting the need for focused educational programs. This analysis helps policymakers prioritize resources by pinpointing specific regions with unique needs, such as improving literacy in Cluster 4 or enhancing employment opportunities in high-literacy clusters.
- **Why K-Means Clustering?**: K-means clustering is chosen because it is an effective method for partitioning data into groups based on similarity across multiple variables. It is particularly useful for demographic studies, as it can reveal hidden patterns (e.g., low-literacy, high-workforce areas) that aggregate statistics might overlook.



Initial Cluster Centers

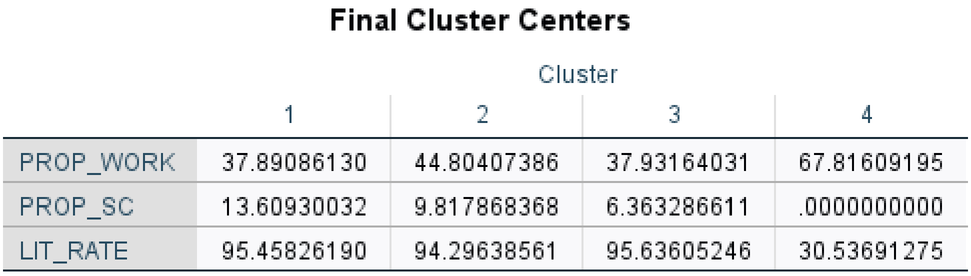
	Cluster			
	1	2	3	4
PROP_WORK	39.79469157	47.91973620	33.88278388	67.81609195
PROP_SC	26.39415518	11.98344208	1.465201465	.0000000000
LIT_RATE	95.49031477	90.25058731	97.98387097	30.53691275



Iteration History[a]

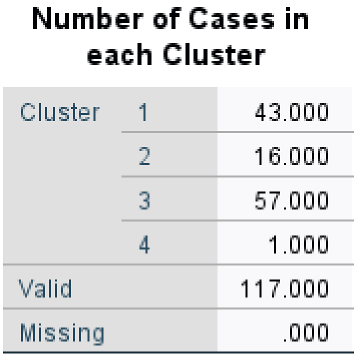
Iteration	Change in Cluster Centers 1	2	3	4
1	8.694	7.575	7.261	.000
2	.987	.398	.148	.000
3	.589	.000	.164	.000
4	.222	.127	.000	.000
5	.482	.253	.083	.000
6	.437	.417	.000	.000
7	.511	1.087	.122	.000
8	.642	.996	.179	.000
9	.423	.279	.269	.000
10	.171	.000	.130	.000

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum absolute coordinate change for any center is .171. The current iteration is 10. The minimum distance between initial centers is 17.353.



Final Cluster Centers

	Cluster 1	2	3	4
PROP_WORK	37.89086130	44.80407386	37.93164031	67.81609195
PROP_SC	13.60930032	9.817868368	6.363286611	.0000000000
LIT_RATE	95.45826190	94.29638561	95.63605246	30.53691275



Number of Cases in each Cluster

Cluster	1	43.000
	2	16.000
	3	57.000
	4	1.000
Valid		117.000
Missing		.000

Cluster Analysis Conclusion:

The cluster analysis results indicate that the demographic and socioeconomic characteristics of Ernakulam district can be grouped into four distinct clusters. Key observations include:

- **Stability in Cluster 4:** The values for PROP_WORK, PROP_SC, and LIT_RATE remained unchanged in Cluster 4, suggesting that this cluster represents a distinct group with consistent characteristics across iterations. This may indicate a highly differentiated segment of the population, possibly a rural or less-developed area with lower literacy rates (30.54%).
- **Changes in Work and Socioeconomic Proportions:** Clusters 1, 2, and 3 saw slight shifts in the proportions of working populations (PROP_WORK) and scheduled caste representation (PROP_SC). These changes reflect the iterative refinement process in k-means clustering, aiming to better separate the groups.
- **Literacy Rate Variation:** The literacy rates across clusters show distinct groupings, with Cluster 4 having significantly lower literacy (30.54%) compared to Clusters 1, 2, and 3, which exhibit much higher literacy rates (~95%). This suggests that certain regions have a strong correlation between employment and education.
- **Policy Implications:** The final clusters suggest varying levels of literacy and workforce participation across Ernakulam. Policymakers can leverage these insights to target specific areas for educational and employment interventions, particularly in low-literacy clusters.

Overall, the clustering confirms that Ernakulam district exhibits socioeconomic diversity, with distinct patterns in work participation, literacy, and caste representation.

5.4 T-Test

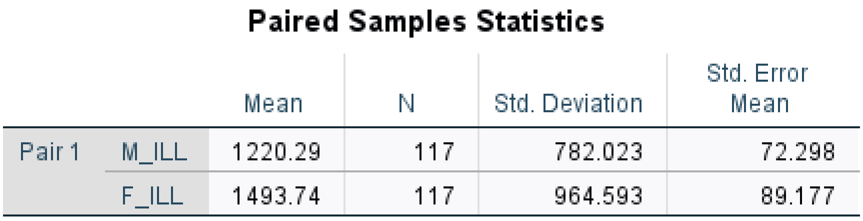
"This section uses paired t-tests to compare literacy and illiteracy rates between males and females across towns in Ernakulam district, with the aim of quantifying gender disparities in educational attainment. Understanding these differences is crucial for assessing gender equity in education, a cornerstone of Kerala's development model, and for identifying potential barriers to female participation in the workforce, which is explored in later sections. By focusing on paired observations (male and female rates from the same towns), this analysis ensures a direct comparison while controlling for town-specific variations, making the paired t-test an appropriate method for detecting significant gender differences in literacy and illiteracy."

- **Purpose**: The paired t-test is used to compare literacy and illiteracy rates between males and females in the same towns, focusing on gender disparities in educational attainment.
- **Reason/Aim**: The aim is to quantify the extent of gender differences in literacy and illiteracy, which is critical for assessing gender equity in education—a key aspect of

Kerala’s development model. The significant difference (2.66%) indicates that, despite Kerala’s high overall literacy, females lag slightly behind males, which has implications for gender-focused educational policies. This analysis also provides a foundation for understanding gender disparities in workforce participation (explored later), as literacy is a key determinant of employment.

- **Why Paired T-Test?**: A paired t-test is chosen because the data involves paired observations (male and female rates from the same towns), allowing for a direct comparison while accounting for town-specific variations. This method is appropriate for detecting small but significant differences in a paired setting.

5.4.1 Gender Disparity Illiteracy T-Test



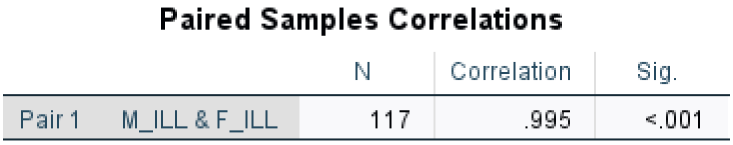
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	M_ILL	1220.29	117	782.023	72.298
	F_ILL	1493.74	117	964.593	89.177

1. **Mean Comparison:**
   - The mean value of **F_ILL** (1493.74) is higher than the mean value of **M_ILL** (1220.29). This indicates that, on average, the female illiterate population (F_ILL) is higher than the male illiterate population (M_ILL).
2. **Variation and Consistency:**
   - The standard deviation for **F_ILL** (964.593) is higher than that for **M_ILL** (782.023), indicating that the illiteracy rates among females show more variability compared to males.
   - The standard error mean for **F_ILL** (89.177) is also higher than that for **M_ILL** (72.298), suggesting that the estimated mean for female illiteracy has a higher variability.

Implications:

1. **Policy Focus:**
   - The higher mean illiteracy rate among females suggests that there may be a need for targeted policies and interventions to address female illiteracy more effectively.
2. **Resource Allocation:**
   - Resources for literacy programs might be more effectively utilized if they are directed towards female populations, given the higher mean illiteracy rate and variability.
3. **Further Research:**
   - Investigate the underlying factors contributing to higher female illiteracy rates and variability to design more effective interventions.



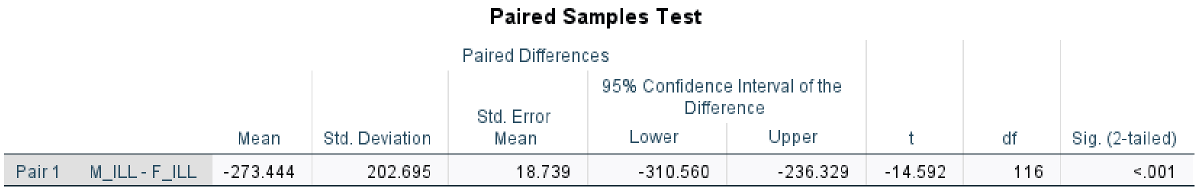
Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	M_ILL & F_ILL	117	.995	<.001

1. **Strong Positive Correlation:**
   - The Pearson correlation coefficient between **M_ILL** and **F_ILL** is .995. This indicates an extremely strong positive correlation between the male illiteracy rate and the female illiteracy rate.
2. **Statistical Significance:**
   - The significance level (Sig.) is less than .001. This means that the correlation is statistically significant at the 0.001 level, indicating a very strong and reliable relationship between the two variables.

Overall Implications:

1. **Interdependence:**
    - The strong positive correlation suggests that as the male illiteracy rate increases, the female illiteracy rate also increases, and vice versa. This indicates that factors influencing illiteracy rates affect both genders similarly.
2. **Policy Focus:**
    - Interventions aimed at reducing illiteracy rates should consider addressing both male and female illiteracy together, as they are closely linked.
3. **Resource Allocation:**
    - Resources for literacy programs might be more effective if they are directed towards both male and female populations simultaneously, given the strong interdependence.
4. **Further Research:**
    - Investigate the underlying factors contributing to the strong correlation between male and female illiteracy rates to design more effective literacy interventions.



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	M_ILL - F_ILL	-273.444	202.695	18.739	-310.560	-236.329	-14.592	116	<.001

The table titled "Paired Samples Test" compares the differences between two paired samples, **M_ILL** (Male Illiteracy Rate) and **F_ILL** (Female Illiteracy Rate). Here are the key statistics:

1. **Mean Difference:**
    - **Mean:** -273.444
    - This indicates that, on average, the male illiteracy rate is 273.444 units lower than the female illiteracy rate.
2. **Standard Deviation:**

- **Std. Deviation:** 202.695
- This shows the amount of variation in the differences between the male and female illiteracy rates.

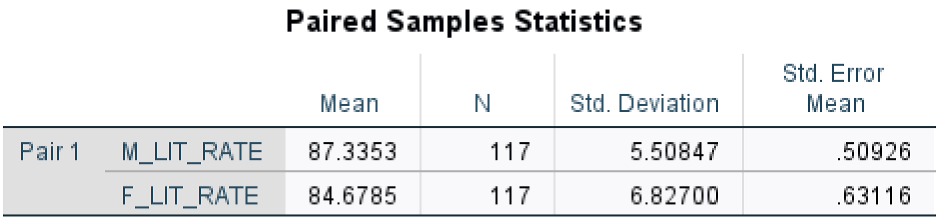
3. **Standard Error Mean:**
    - **Std. Error Mean:** 18.739
    - This is the standard error of the mean difference, indicating the precision of the mean difference estimate.
4. **95% Confidence Interval of the Difference:**
    - **Lower Bound:** -310.560
    - **Upper Bound:** -236.329
    - This interval indicates that we are 95% confident that the true mean difference between the male and female illiteracy rates lies between -310.560 and -236.329.
5. **t-value:**
    - **t:** -14.592
    - This is the t-statistic for the paired sample test, used to determine the significance of the mean difference.
6. **Degrees of Freedom (df):**
    - **df:** 116
    - The degrees of freedom for the test, which is the number of paired samples minus one.
7. **Significance (Sig. 2-tailed):**
    - **Sig. (2-tailed):** < .001
    - This indicates that the difference in illiteracy rates between males and females is highly statistically significant. Since the p-value is less than .001, we reject the null hypothesis that there is no difference in illiteracy rates between males and females.

Overall Implications:

1. **Significant Difference:**
    - There is a highly significant difference between male and female illiteracy rates, with females having higher illiteracy rates on average.
2. **Policy Focus:**
    - This significant difference suggests a need for targeted policies and interventions to reduce female illiteracy rates, addressing the underlying factors contributing to this disparity.
3. **Resource Allocation:**
    - Resources for literacy programs might be more effectively utilized if they are directed towards female populations, given the higher mean illiteracy rate and significant difference.
4. **Further Research:**
    - Investigate the socio-economic, cultural, and educational factors contributing to higher female illiteracy rates to design more effective interventions.

5.4.2 Gender Disparity Literacy T-Test

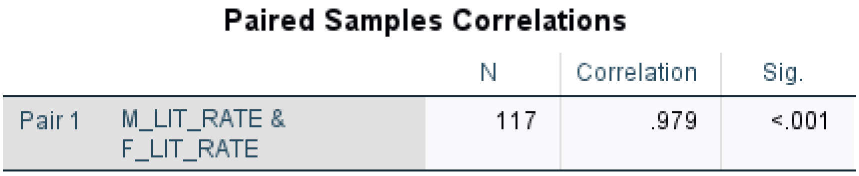
Male Literacy Rate and **Female Literacy Rate** for each town in Kerala, the best approach is a **paired t-test**, because the literacy rates for males and females are from the same towns (paired observations).



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	M_LIT_RATE	87.3353	117	5.50847	.50926
	F_LIT_RATE	84.6785	117	6.82700	.63116

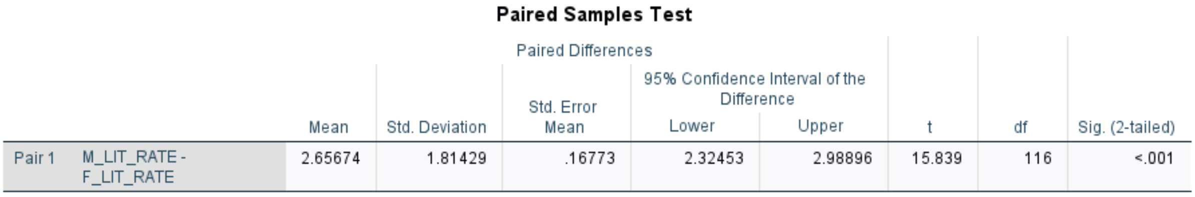
- On average, male literacy rates are higher than female literacy rates across towns.
- There is more variability in female literacy rates than male literacy rates.



Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	M_LIT_RATE & F_LIT_RATE	117	.979	<.001

- Male and female literacy rates tend to move together across towns.
- Even though there is a gender gap in literacy (as seen in the previous table), male and female literacy rates are **highly dependent on each other**.



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	M_LIT_RATE - F_LIT_RATE	2.65674	1.81429	.16773	2.32453	2.98896	15.839	116	<.001

Key Insights:

1. **Mean Difference (2.65674):**
   - On average, male literacy rates are **2.66 percentage points higher** than female literacy rates across towns.
   - This confirms a **gender disparity in literacy** favoring males.
2. **Standard Deviation of the Difference (1.81429):**
   - The variability in the literacy rate differences across towns.
3. **Standard Error Mean (0.16773):**
   - Indicates the precision of the estimated mean difference.
4. **95% Confidence Interval (2.32453 to 2.98896):**
   - We are **95% confident** that the true mean difference in literacy rates (Male - Female) lies between **2.32% and 2.99%**.
   - Since this interval does **not** include **0**, it confirms a significant difference.
5. **t-Value (15.839):**
   - A high t-value indicates a strong effect of gender on literacy rates.
6. **Degrees of Freedom (df = 116):**
   - Since we have 117 towns, the degrees of freedom are **N - 1 = 116**.
7. **Significance (p-value = 0.000):**

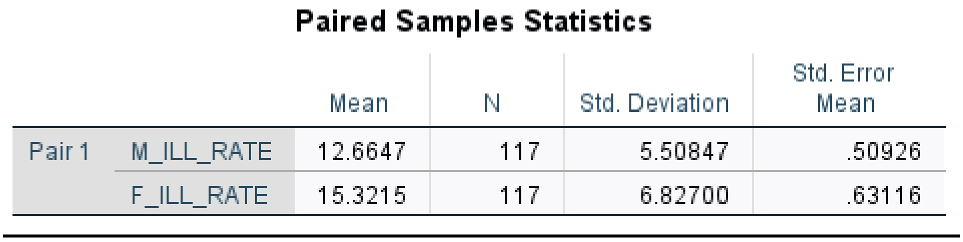
- Since **p < 0.05**, the difference is **statistically significant**.
- This means that the higher male literacy rate is **not due to random chance**.

Conclusion

There is a statistically significant gender disparity in literacy rates in Kerala's towns, with males having a higher literacy rate than females (by about 2.66%).

5.4.3 Gender Disparity Illiteracy T-Test

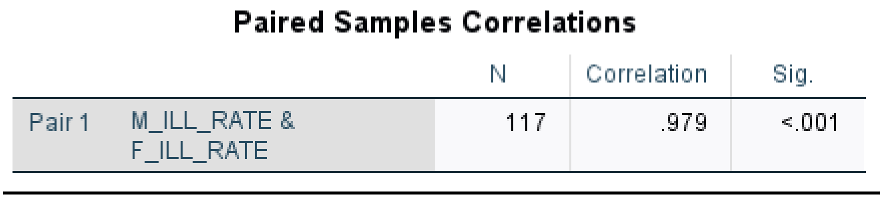
Male Literacy Rate and **Female Literacy Rate** for each town in Kerala, the best approach is a **paired t-test**, because the illiteracy rates for males and females are from the same towns (paired observations).



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	M_ILL_RATE	12.6647	117	5.50847	.50926
	F_ILL_RATE	15.3215	117	6.82700	.63116

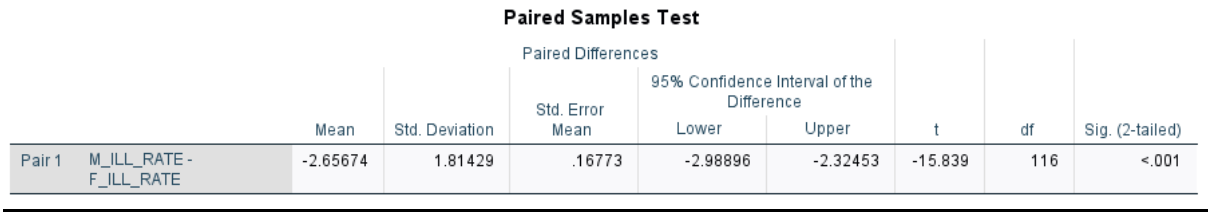
- **Female illiteracy is higher than male illiteracy across towns.**
- **There is more variability in female illiteracy rates than male illiteracy rates.**



Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	M_ILL_RATE & F_ILL_RATE	117	.979	<.001

- Similar to literacy rates, male and female illiteracy rates are highly dependent on each other.



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	M_ILL_RATE - F_ILL_RATE	-2.65674	1.81429	.16773	-2.98896	-2.32453	-15.839	116	<.001

Key Insights:

1. **Mean Difference (-2.65674):**
   - The difference is **negative**, meaning **female illiteracy is higher than male illiteracy** by **2.66 percentage points** on average.
   - This confirms a **gender disparity in illiteracy**, with more females being illiterate compared to males.
2. **Standard Deviation of the Difference (1.81429):**
   - Indicates how much the difference in illiteracy rates varies across towns.
3. **Standard Error Mean (0.16773):**
   - A small value, meaning the estimate of the mean difference is precise.
4. **95% Confidence Interval (-2.98896 to -2.32453):**
   - We are **95% confident** that the true mean difference (Male - Female) lies between **-2.99% and -2.32%**.
   - Since this interval does **not include 0**, it confirms a **significant difference** in illiteracy rates between genders.
5. **t-Value (-15.839):**
   - A very large negative t-value, indicating a strong difference between male and female illiteracy rates.
6. **Degrees of Freedom (df = 116):**
   - Since we have 117 towns, the degrees of freedom are **N - 1 = 116**.
7. **Significance (p-value = 0.000):**
   - Since **p < 0.05**, the difference is **statistically significant**.

- This means that the higher female illiteracy rate is **not due to random chance**.

Conclusion

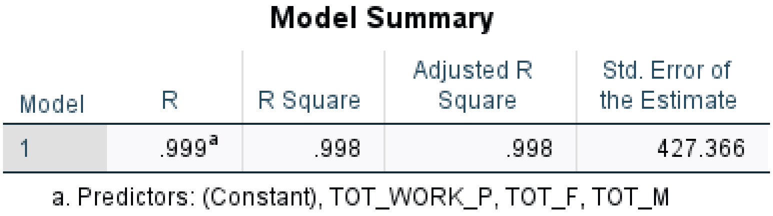
There is a statistically significant gender disparity in illiteracy rates in Kerala's towns, with females having a higher illiteracy rate than males (by about 2.66%).

5.5 Regression

“Regression analysis is employed in this section to model the relationships between literacy rate and workforce participation (dependent variables) and various predictors, such as population, SC/ST proportions, and agricultural labourers, with the aim of identifying key drivers of these outcomes in Ernakulam district. By quantifying the impact of factors like literacy rate, sex ratio, and marginalized groups on workforce participation, this analysis provides actionable insights for policy formulation, such as enhancing education to boost employment or targeting specific sectors (e.g., agriculture) for economic development. Multiple linear regression is selected for its ability to examine multiple predictors simultaneously, offering a comprehensive understanding of their combined effects on the dependent variables, which aligns with the complexity of the census data.”

- **Purpose**: Regression analysis is conducted to model the relationships between dependent variables (literacy rate, workforce participation) and multiple independent variables (e.g., population, SC/ST proportions, agricultural labourers), identifying which factors significantly influence these outcomes.
- **Reason/Aim**: The aim is to understand the drivers of literacy and workforce participation in Ernakulam, providing actionable insights for policy formulation. For example, finding that literacy rate and sex ratio positively affect workforce participation (B = 0.0401) suggests that improving education and gender balance can boost employment, which is critical for economic development. Similarly, the positive effect of female population on literacy highlights the potential for increasing female workforce participation through education. This analysis also helps quantify the impact of marginalized groups (e.g., ST population, B = 0.528, $p < 0.001$) and agricultural sectors on employment, informing targeted interventions.
- **Why Multiple Linear Regression?**: Multiple linear regression is chosen because it allows for the simultaneous examination of multiple predictors, providing a comprehensive view of their combined effects on the dependent variable. It is suitable for your dataset, which includes continuous variables (e.g., literacy rate, workforce participation) and multiple potential influencers.

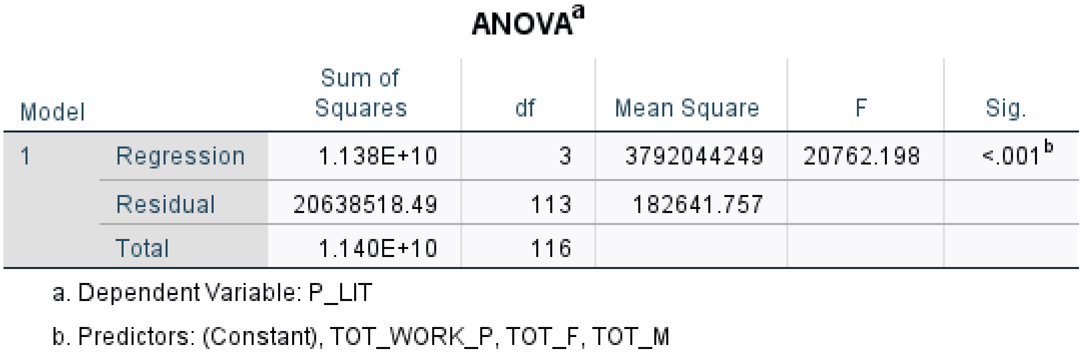
5.5.1 Regression Predict Literacy:(Dependent Variable P_LIT Independent Variables: TOT_P, TOT_M, TOT_F, TOT_WORK_P)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.998	.998	427.366

a. Predictors: (Constant), TOT_WORK_P, TOT_F, TOT_M



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.138E+10	3	3792044249	20762.198	<.001[b]
	Residual	20638518.49	113	182641.757		
	Total	1.140E+10	116			

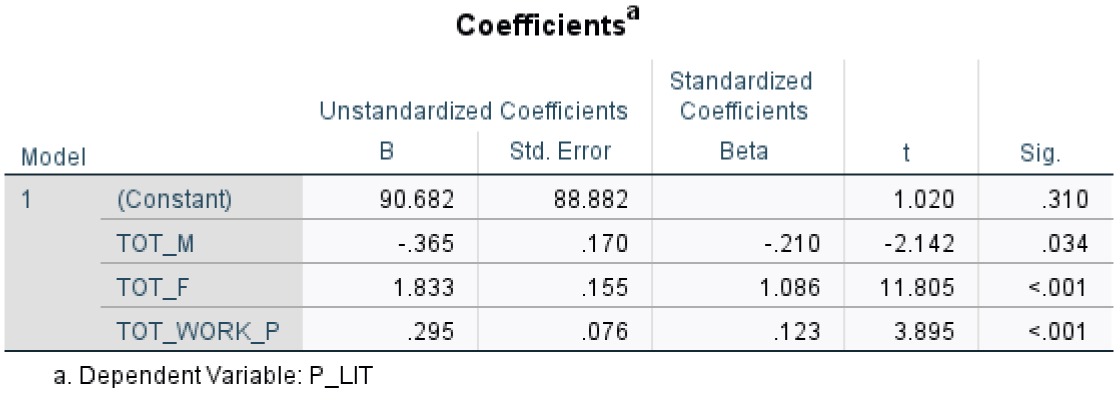
a. Dependent Variable: P_LIT

b. Predictors: (Constant), TOT_WORK_P, TOT_F, TOT_M

Conclusion

The regression model is statistically significant, as indicated by the Significance (Sig.) value of less than .001. This means that the predictors (TOT_WORK_P, TOT_F, TOT_M) collectively have a significant effect on the dependent variable (P_LIT). The high F-statistic (20762.198) further supports the significance of the model.

- **Significant Predictors**: TOT_WORK_P (total working population), TOT_F (total female population), and TOT_M (total male population) are significant predictors of P_LIT (presumably literacy rate).
- **Policy Formulation**: Policies aimed at improving the literacy rate should consider these predictors collectively, as they have a significant impact on the literacy rate.
- **Resource Allocation**: Resources and efforts can be directed towards improving the factors represented by these predictors to positively influence the literacy rate.



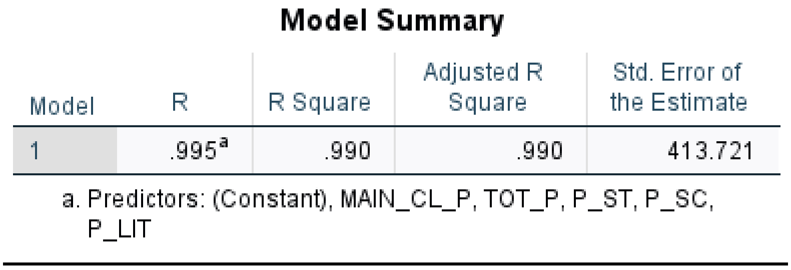
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	90.682	88.882		1.020	.310
	TOT_M	-.365	.170	-.210	-2.142	.034
	TOT_F	1.833	.155	1.086	11.805	<.001
	TOT_WORK_P	.295	.076	.123	3.895	<.001

a. Dependent Variable: P_LIT

- **Negative Influence of Male Population:** A higher male population tends to decrease literacy rates, which may warrant investigation into the underlying causes, such as male employment patterns or societal expectations.
- **Positive Influence of Female and Working Population:** Increasing both the female and working populations significantly boosts literacy rates, highlighting the importance of gender inclusivity and employment in promoting literacy.

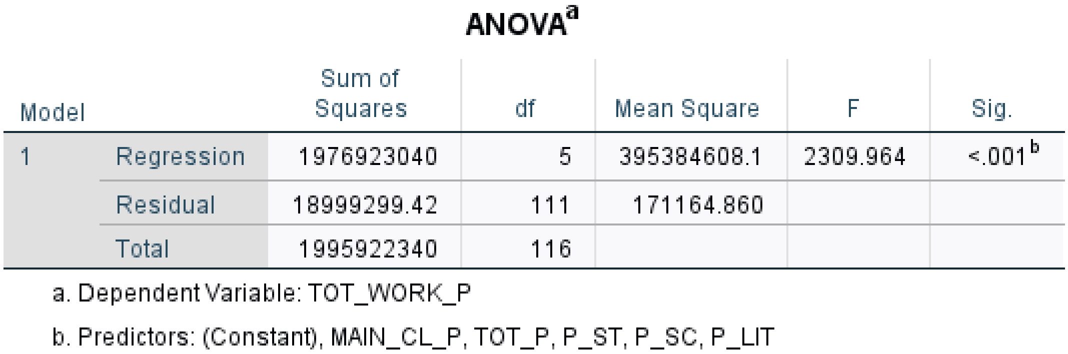
5.5.2 Regression Predict Work Participation



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995[a]	.990	.990	413.721

a. Predictors: (Constant), MAIN_CL_P, TOT_P, P_ST, P_SC, P_LIT



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1976923040	5	395384608.1	2309.964	<.001[b]
	Residual	18999299.42	111	171164.860		
	Total	1995922340	116			

a. Dependent Variable: TOT_WORK_P

b. Predictors: (Constant), MAIN_CL_P, TOT_P, P_ST, P_SC, P_LIT

Inference from the Regression Table:

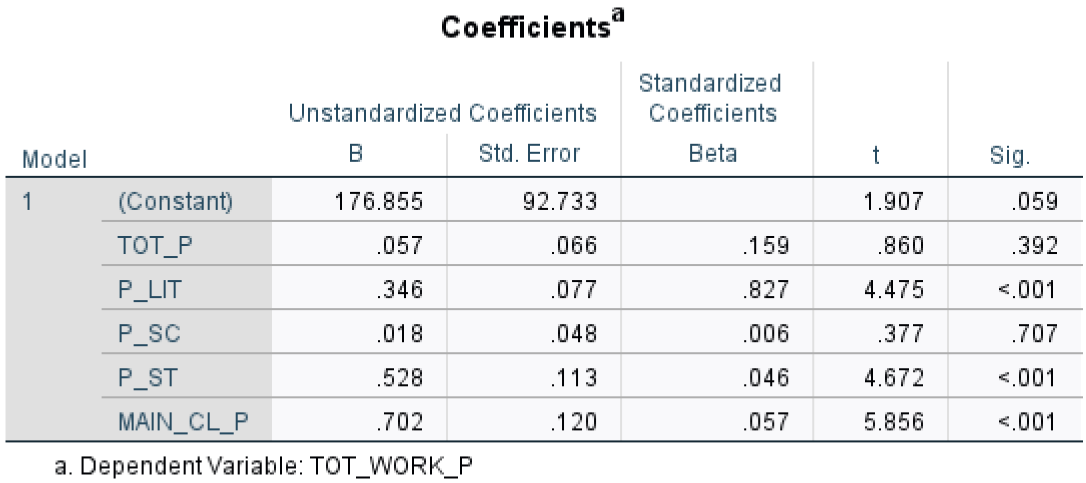
The regression results provide insights into the factors influencing the dependent variable (TOT_WORK_P), which is the total working population in this context. Here are the key points:

1. **Model Significance:**
   - The F-value is 2309.964, and the significance level (Sig.) is less than .001. This indicates that the overall regression model is statistically significant, meaning **the predictors collectively have a significant impact on the dependent variable**.
2. **Predictors:**
   - The predictors in the model are MAIN_CL_P (main cultivators population percentage), TOT_P (total population), P_ST (scheduled tribes population percentage), P_SC (scheduled castes population percentage), and P_LIT (literacy rate).
   - These predictors have a significant influence on the total working population, as indicated by the model's significance.
3. **Regression Coefficients:**
   - The coefficients for each predictor would provide specific insights into the direction and magnitude of their influence on the dependent variable. However, these coefficients are not provided in the current table. Typically, you would look at the unstandardized coefficients (B) and their significance levels to understand the impact of each predictor.

Implications:

- **Policy Formulation:** Understanding the significant predictors of the working population can help policymakers design targeted interventions to improve employment rates. For instance, improving literacy rates or addressing the needs of scheduled tribes and castes can have a positive impact on employment.

- **Resource Allocation:** Resources can be allocated more efficiently by focusing on the significant predictors identified in the regression model.
- **Further Research:** This analysis provides a foundation for further research into the specific factors affecting the working population, allowing for more nuanced studies and interventions.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	176.855	92.733		1.907	.059
	TOT_P	.057	.066	.159	.860	.392
	P_LIT	.346	.077	.827	4.475	<.001
	P_SC	.018	.048	.006	.377	.707
	P_ST	.528	.113	.046	4.672	<.001
	MAIN_CL_P	.702	.120	.057	5.856	<.001

a. Dependent Variable: TOT_WORK_P

- **Constant:**
  - The constant term has a B value of 176.855 with a standard error of 92.733, a t-value of 1.907, and a significance level (Sig.) of .059.
  - **Inference:** The constant term is not statistically significant (p-value > .05), indicating that the intercept of the regression line is not significantly different from zero.
- **TOT_P (Total Population):**
  - The B value is .057, with a standard error of .066, a Beta of .159, a t-value of .860, and a significance level of .392.
  - **Inference:** The total population does not have a statistically significant effect on the dependent variable (TOT_WORK_P), as the p-value is greater than .05.
- **P_LIT (Literacy Rate):**
  - The B value is .346, with a standard error of .077, a Beta of .827, a t-value of 4.475, and a significance level of less than .001.

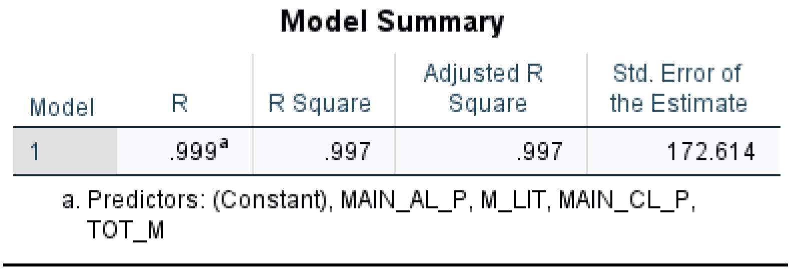
    - **Inference:** The literacy rate has a positive and statistically significant effect on the total working population. This suggests that higher literacy rates are associated with an increase in the working population.
- **P_SC (Scheduled Castes Population Percentage):**
    - The B value is .018, with a standard error of .048, a Beta of .006, a t-value of .377, and a significance level of .707.
    - **Inference:** The scheduled castes population percentage does not have a statistically significant effect on the dependent variable, as the p-value is greater than .05.
- **P_ST (Scheduled Tribes Population Percentage):**
    - The B value is .528, with a standard error of .113, a Beta of .046, a t-value of 4.672, and a significance level of less than .001.
    - **Inference:** The scheduled tribes population percentage has a positive and statistically significant effect on the total working population. This indicates that an increase in the scheduled tribes population is associated with an increase in the working population.
- **MAIN_CL_P (Main Cultivators Population Percentage):**
    - The B value is .702, with a standard error of .120, a Beta of .057, a t-value of 5.856, and a significance level of less than .001.
    - **Inference:** The main cultivators population percentage has a positive and statistically significant effect on the total working population. This suggests that a higher percentage of main cultivators is associated with an increase in the working population.

Overall Implications:

- **Positive Influences:** The literacy rate, scheduled tribes population percentage, and main cultivators population percentage all have significant positive effects on the total working population. This highlights the importance of these factors in boosting employment rates.

- **Insignificant Factors:** The total population and scheduled castes population percentage do not have significant effects on the working population, suggesting that other factors may play a more critical role in influencing employment.

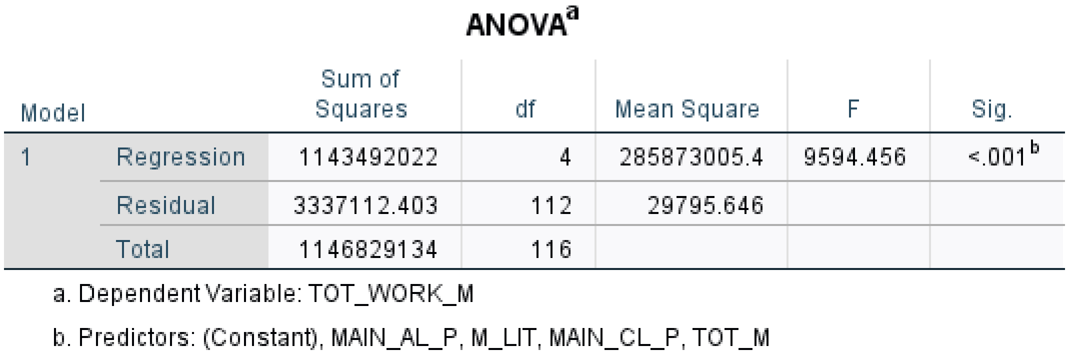
5.5.3 Regression Predict Gender-Based Workforce Participation(Male)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.997	.997	172.614

a. Predictors: (Constant), MAIN_AL_P, M_LIT, MAIN_CL_P, TOT_M



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1143492022	4	285873005.4	9594.456	<.001[b]
	Residual	3337112.403	112	29795.646		
	Total	1146829134	116			

a. Dependent Variable: TOT_WORK_M

b. Predictors: (Constant), MAIN_AL_P, M_LIT, MAIN_CL_P, TOT_M

Model Significance:

- The F-value is 9594.456 with a significance level (Sig.) of less than .001. This indicates that the overall regression model is highly statistically significant, meaning that the predictors collectively have a significant impact on the dependent variable (TOT_WORK_M).

3. **Predictors:** The predictors in the model are MAIN_AL_P (main agricultural laborers population percentage), M_LIT (male literacy rate), MAIN_CL_P (main cultivators population percentage), and TOT_M (total male population).

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	33.671	38.102		.884	.379
	TOT_M	.322	.068	.584	4.724	.000
	M_LIT	.263	.078	.414	3.360	.001
	MAIN_CL_P	.040	.103	.004	.388	.699
	MAIN_AL_P	.172	.098	.020	1.767	.080

a. Dependent Variable: TOT_WORK_M

Regression Coefficients:

Model Significance:

- **F-value**: 9594.456
- **Significance Level (Sig.)**: < .001

The overall regression model is highly statistically significant, meaning that the predictors collectively have a significant impact on the dependent variable (TOT_WORK_M - Total Working Male Population).

Regression Coefficients:

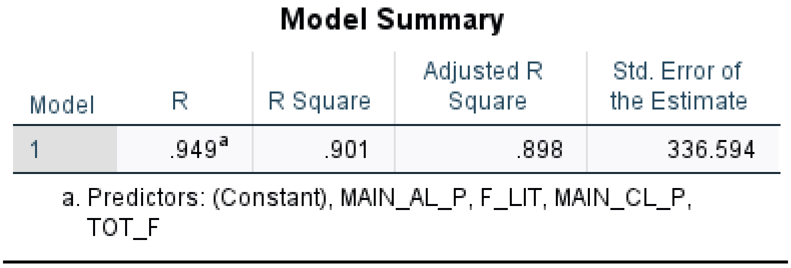
- **Constant:**
  - B = 33.671
  - Std. Error = 38.102
  - t-value = .884
  - Sig. = .379
  - **Inference:** The constant term is not statistically significant (p-value > .05), indicating that the intercept of the regression line is not significantly different from zero.
- **TOT_M (Total Male Population):**
  - B = .322

    - Std. Error = .068
    - t-value = 4.724
    - Sig. < .001
    - **Inference:** The total male population has a positive and statistically significant effect on the total working male population. This suggests that an increase in the total male population is associated with an increase in the working male population.
- **M_LIT (Male Literacy Rate):**
    - B = .263
    - Std. Error = .078
    - t-value = 3.360
    - Sig. = .001
    - **Inference:** The male literacy rate has a positive and statistically significant effect on the total working male population. Higher male literacy rates are associated with an increase in the working male population.
- **MAIN_CL_P (Main Cultivators Population Percentage):**
    - B = .040
    - Std. Error = .103
    - t-value = .388
    - Sig. = .699
    - **Inference:** The main cultivators population percentage does not have a statistically significant effect on the total working male population, as the p-value is greater than .05.
- **MAIN_AL_P (Main Agricultural Laborers Population Percentage):**
    - B = .172
    - Std. Error = .098
    - t-value = 1.767
    - Sig. = .080
    - **Inference:** The main agricultural laborers population percentage is close to being statistically significant (p-value = .080), suggesting a potential positive effect on the total working male population, but this effect is not definitive.

Overall Implications:

- **Positive Influences:** The total male population and male literacy rate have significant positive effects on the working male population.
- **Insignificant Factors:** The main cultivators population percentage does not have a significant effect, and the main agricultural laborers population percentage is not definitively significant.
- **Policy Implications:** Focus on improving male literacy rates and consider factors influencing the male population to enhance the working male population.

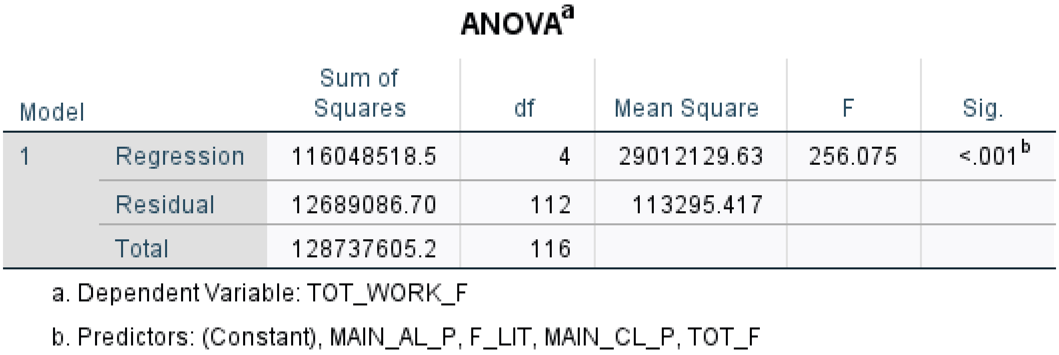
5.5.4 Regression Predict Gender-Based Workforce Participation(Female)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.949[a]	.901	.898	336.594

a. Predictors: (Constant), MAIN_AL_P, F_LIT, MAIN_CL_P, TOT_F



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	116048518.5	4	29012129.63	256.075	<.001[b]
	Residual	12689086.70	112	113295.417		
	Total	128737605.2	116			

a. Dependent Variable: TOT_WORK_F

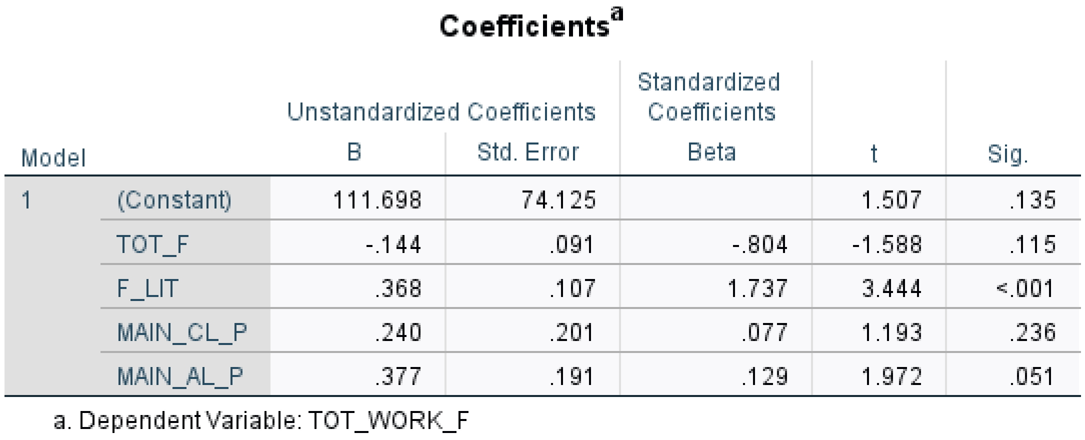
b. Predictors: (Constant), MAIN_AL_P, F_LIT, MAIN_CL_P, TOT_F

Based on the regression table provided, we can draw the following conclusions:

1. Model Significance:

- **F-value**: 256.075
- **Significance Level (Sig.)**: < .001

The overall regression model is highly statistically significant, meaning that the predictors collectively have a significant impact on the dependent variable (TOT_WORK_F - Total Working Female Population).



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	111.698	74.125		1.507	.135
	TOT_F	-.144	.091	-.804	-1.588	.115
	F_LIT	.368	.107	1.737	3.444	<.001
	MAIN_CL_P	.240	.201	.077	1.193	.236
	MAIN_AL_P	.377	.191	.129	1.972	.051

a. Dependent Variable: TOT_WORK_F

The table shows the results of a regression analysis for the dependent variable **TOT_WORK_F** (Total Working Female Population) and the predictors **TOT_F** (Total Female Population), **F_LIT** (Female Literacy Rate), **MAIN_CL_P** (Main Cultivators Population Percentage), and **MAIN_AL_P** (Main Agricultural Laborers Population Percentage).

Key Findings:

1. **Constant:**
   - B = 111.698
   - Std. Error = 74.125
   - t-value = 1.507
   - Sig. = .135
   - **Inference:** The constant term is not statistically significant (p-value > .05), indicating that the intercept of the regression line is not significantly different from zero.
2. **TOT_F (Total Female Population):**
   - B = -.144
   - Std. Error = .091

- Beta = -.804
- t-value = -1.588
- Sig. = .115
- **Inference:** The total female population does not have a statistically significant effect on the total working female population (p-value > .05).

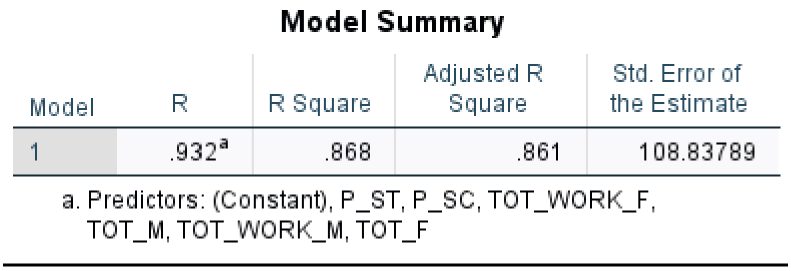
3. **F_LIT (Female Literacy Rate):**
   - B = .368
   - Std. Error = .107
   - Beta = 1.737
   - t-value = 3.444
   - Sig. < .001
   - **Inference:** The female literacy rate has a positive and statistically significant effect on the total working female population. This suggests that higher female literacy rates are associated with an increase in the working female population.
4. **MAIN_CL_P (Main Cultivators Population Percentage):**
   - B = .240
   - Std. Error = .201
   - Beta = .077
   - t-value = 1.193
   - Sig. = .236
   - **Inference:** The main cultivators population percentage does not have a statistically significant effect on the total working female population (p-value > .05).
5. **MAIN_AL_P (Main Agricultural Laborers Population Percentage):**
   - B = .377
   - Std. Error = .191
   - Beta = .129
   - t-value = 1.972
   - Sig. = .051

- **Inference:** The main agricultural laborers population percentage is close to being statistically significant (p-value = .051), suggesting a potential positive effect on the total working female population, but this effect is not definitive.

Overall Implications:

- **Significant Factors:** Female literacy rate is a significant predictor of the working female population, highlighting the importance of improving literacy rates among women to boost employment.
- **Insignificant Factors:** The total female population and main cultivators population percentage do not have significant effects on the working female population.
- **Potential Influence:** The main agricultural laborers population percentage is borderline significant, indicating a possible positive influence on employment among women in agriculture.

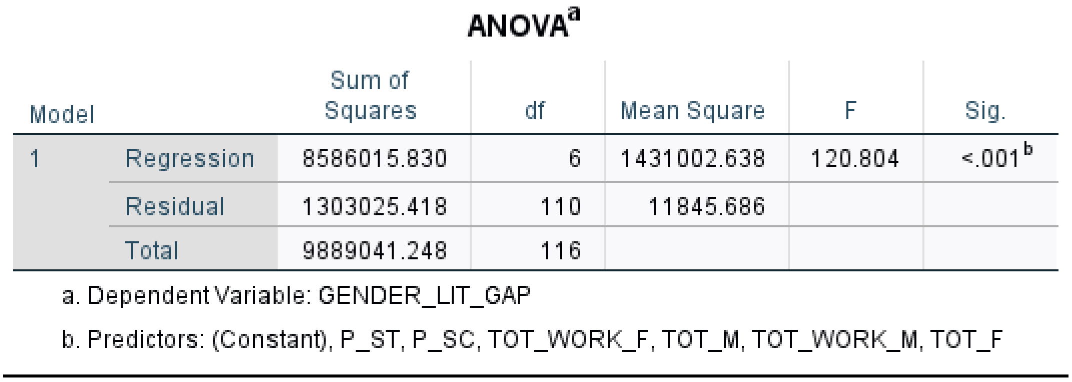
5.5.5 Regression Predict Gender Literacy Gap



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.932[a]	.868	.861	108.83789

a. Predictors: (Constant), P_ST, P_SC, TOT_WORK_F, TOT_M, TOT_WORK_M, TOT_F



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8586015.830	6	1431002.638	120.804	<.001[b]
	Residual	1303025.418	110	11845.686		
	Total	9889041.248	116			

a. Dependent Variable: GENDER_LIT_GAP

b. Predictors: (Constant), P_ST, P_SC, TOT_WORK_F, TOT_M, TOT_WORK_M, TOT_F

The ANOVA (Analysis of Variance) results for the dependent variable **GENDER_LIT_GAP** with predictors **P_ST** (Percentage of Scheduled Tribes), **P_SC** (Percentage of Scheduled Castes),

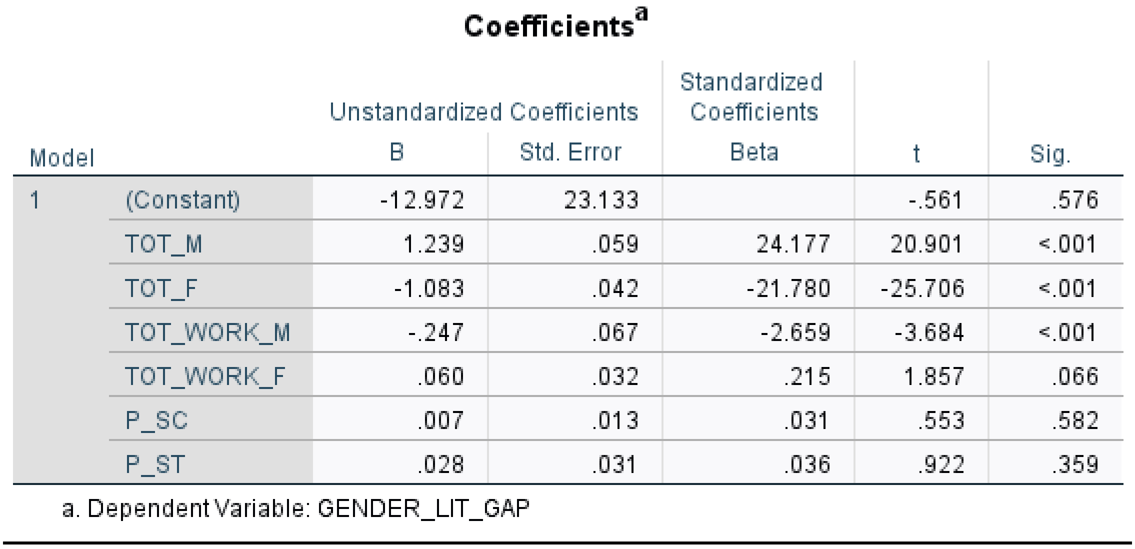
TOT_WORK_F (Total Working Female Population), **TOT_M** (Total Male Population), **TOT_WORK_M** (Total Working Male Population), and **TOT_F** (Total Female Population) indicate the following:

- **Regression:**
  - **Sum of Squares:** 8586015.830
  - **Degrees of Freedom (df):** 6
  - **Mean Square:** 1431002.638
  - **F-value:** 120.804
  - **Significance (Sig.):** < .001
  - **Inference:** The regression model is highly statistically significant, as indicated by the F-value of 120.804 and the p-value of less than .001. This means that the predictors collectively have a significant effect on the dependent variable (GENDER_LIT_GAP).
- **Residual:**
  - **Sum of Squares:** 1303025.418
  - **Degrees of Freedom (df):** 110
  - **Mean Square:** 11845.686
- **Total:**
  - **Sum of Squares:** 9889041.248
  - **Degrees of Freedom (df):** 116

Overall Implications:

1. **Significance of Predictors:** The predictors (P_ST, P_SC, TOT_WORK_F, TOT_M, TOT_WORK_M, TOT_F) collectively have a significant impact on the gender literacy gap.
2. **Model Fit:** The high F-value and low p-value indicate that the regression model fits the data well and the predictors are important in explaining the variation in the gender literacy gap.
3. **Policy Implications:** Understanding the significant predictors of the gender literacy gap can help policymakers design targeted interventions to address literacy disparities. For instance, focusing on improving working conditions for women,

increasing female literacy rates, and addressing socio-economic factors related to scheduled tribes and castes can positively influence the gender literacy gap.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.972	23.133		-.561	.576
	TOT_M	1.239	.059	24.177	20.901	<.001
	TOT_F	-1.083	.042	-21.780	-25.706	<.001
	TOT_WORK_M	-.247	.067	-2.659	-3.684	<.001
	TOT_WORK_F	.060	.032	.215	1.857	.066
	P_SC	.007	.013	.031	.553	.582
	P_ST	.028	.031	.036	.922	.359

a. Dependent Variable: GENDER_LIT_GAP

The table shows the results of a regression analysis for the dependent variable **GENDER_LIT_GAP** with predictors **TOT_M** (Total Male Population), **TOT_F** (Total Female Population), **TOT_WORK_M** (Total Working Male Population), **TOT_WORK_F** (Total Working Female Population), **P_SC** (Percentage of Scheduled Castes), and **P_ST** (Percentage of Scheduled Tribes).

Key Findings:

1. **Constant:**
   - B = -12.972
   - Std. Error = 23.133
   - t-value = -0.561
   - Sig. = .576
   - **Inference:** The constant term is not statistically significant (p-value > .05), indicating that the intercept of the regression line is not significantly different from zero.
2. **TOT_M (Total Male Population):**
   - B = 1.239
   - Std. Error = 0.059

    - Beta = 24.177
    - t-value = 20.901
    - Sig. < .001
    - **Inference:** The total male population has a positive and statistically significant effect on the gender literacy gap. This suggests that an increase in the total male population is associated with an increase in the gender literacy gap.

3. **TOT_F (Total Female Population):**
    - B = -1.083
    - Std. Error = 0.042
    - Beta = -21.780
    - t-value = -25.706
    - Sig. < .001
    - **Inference:** The total female population has a negative and statistically significant effect on the gender literacy gap. This indicates that an increase in the total female population is associated with a decrease in the gender literacy gap.
4. **TOT_WORK_M (Total Working Male Population):**
    - B = -0.247
    - Std. Error = 0.067
    - Beta = -2.659
    - t-value = -3.684
    - Sig. < .001
    - **Inference:** The total working male population has a negative and statistically significant effect on the gender literacy gap. This suggests that an increase in the working male population is associated with a decrease in the gender literacy gap.
5. **TOT_WORK_F (Total Working Female Population):**
    - B = 0.060
    - Std. Error = 0.032
    - Beta = -0.215
    - t-value = 1.857
    - Sig. = .066

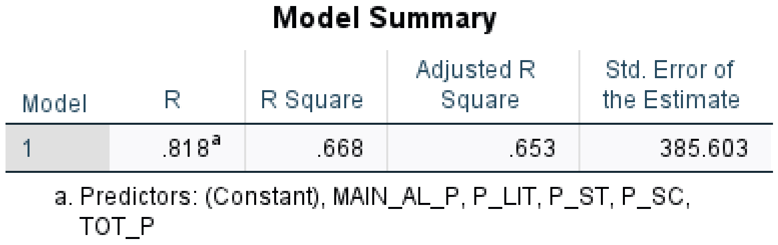
- **Inference:** The total working female population is close to being statistically significant (p-value = .066), indicating a potential positive effect on the gender literacy gap, but this effect is not definitive.

6. **P_SC (Percentage of Scheduled Castes):**
    - B = 0.007
    - Std. Error = 0.013
    - Beta = 0.031
    - t-value = 0.553
    - Sig. = .582
    - **Inference:** The percentage of scheduled castes does not have a statistically significant effect on the gender literacy gap (p-value > .05).
7. **P_ST (Percentage of Scheduled Tribes):**
    - B = 0.028
    - Std. Error = 0.031
    - Beta = 0.036
    - t-value = 0.922
    - Sig. = .359
    - **Inference:** The percentage of scheduled tribes does not have a statistically significant effect on the gender literacy gap (p-value > .05).

Overall Implications:

- **Positive Influences:** The total male population has a significant positive effect on the gender literacy gap, suggesting that higher male populations may widen the gap.
- **Negative Influences:** The total female population and total working male population have significant negative effects on the gender literacy gap, indicating that increasing these populations may help reduce the gap.
- **Potential Influence:** The total working female population is close to being significant, suggesting a possible positive influence on reducing the gap, but this is not definitive.
- **Insignificant Factors:** The percentage of scheduled castes and scheduled tribes do not have significant effects on the gender literacy gap.

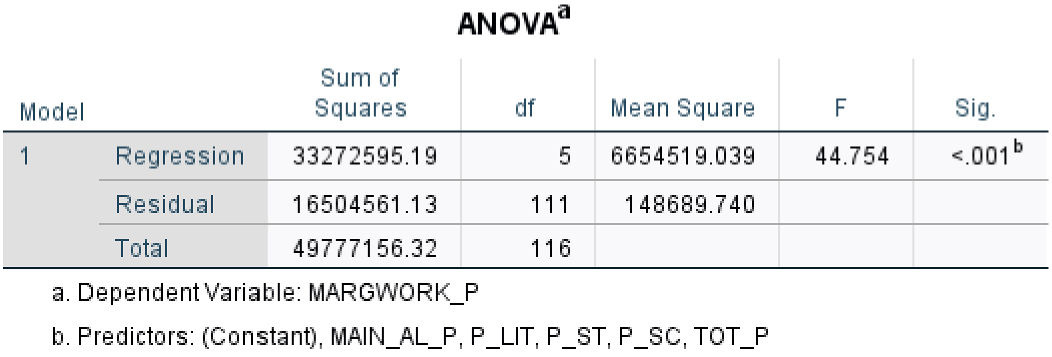
5.5.6 Regression Marginal Workforce Participation



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.818[a]	.668	.653	385.603

a. Predictors: (Constant), MAIN_AL_P, P_LIT, P_ST, P_SC, TOT_P



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33272595.19	5	6654519.039	44.754	<.001[b]
	Residual	16504561.13	111	148689.740		
	Total	49777156.32	116			

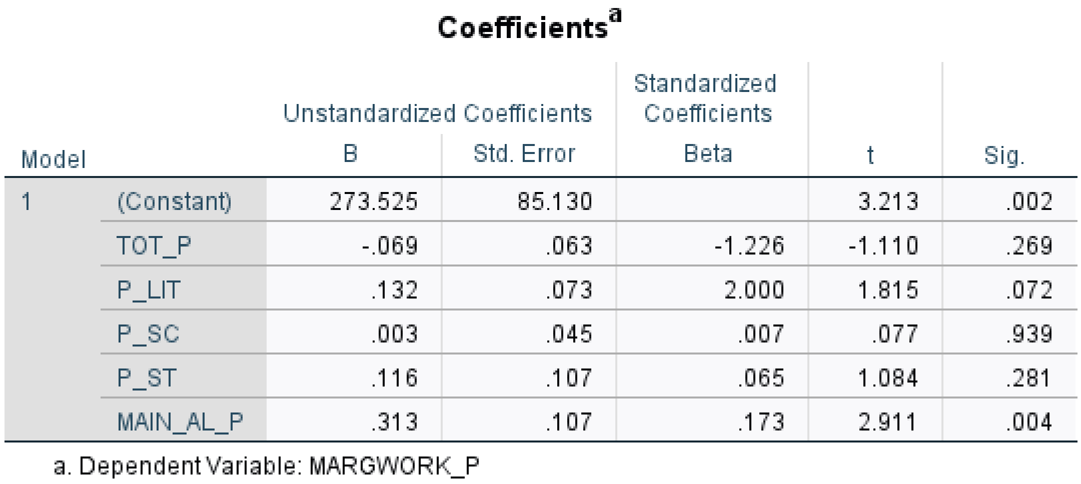
a. Dependent Variable: MARGWORK_P

b. Predictors: (Constant), MAIN_AL_P, P_LIT, P_ST, P_SC, TOT_P

1. **Regression:**
   - **Sum of Squares:** 33,272,595.19
   - **Degrees of Freedom (df):** 5
   - **Mean Square:** 6,654,519.039
   - **F-value:** 44.754
   - **Significance (Sig.):** < .001
   - **Inference:** The regression model is highly statistically significant, as indicated by the F-value of 44.754 and the p-value of less than .001. This means that the predictors collectively have a significant effect on the dependent variable (MARGWORK_P).
2. **Residual:**
   - **Sum of Squares:** 16,504,561.13
   - **Degrees of Freedom (df):** 111
   - **Mean Square:** 148,689.740
3. **Total:**
   - **Sum of Squares:** 49,777,156.32
   - **Degrees of Freedom (df):** 116

Overall Implications:

1. **Significance of Predictors:** The predictors (MAIN_AL_P, P_LIT, P_ST, P_SC, TOT_P) collectively have a significant impact on the marginal workers population percentage.
2. **Model Fit:** The high F-value and low p-value indicate that the regression model fits the data well and the predictors are important in explaining the variation in the marginal workers population percentage.
3. **Policy Implications:** Understanding the significant predictors of the marginal workers population percentage can help policymakers design targeted interventions to address employment issues. For instance, focusing on improving literacy rates, supporting agricultural laborers, and addressing socio-economic factors related to scheduled tribes and castes can positively influence the marginal workers population.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	273.525	85.130		3.213	.002
	TOT_P	-.069	.063	-1.226	-1.110	.269
	P_LIT	.132	.073	2.000	1.815	.072
	P_SC	.003	.045	.007	.077	.939
	P_ST	.116	.107	.065	1.084	.281
	MAIN_AL_P	.313	.107	.173	2.911	.004

a. Dependent Variable: MARGWORK_P

The table shows the results of a regression analysis for the dependent variable **MARGWORK_P** (Marginal Workers Population Percentage) with predictors **TOT_P** (Total Population), **P_LIT** (Literacy Rate), **P_SC** (Percentage of Scheduled Castes), **P_ST** (Percentage of Scheduled Tribes), and **MAIN_AL_P** (Main Agricultural Laborers Population Percentage).

Key Findings:

1. **Constant:**
   - B = 273.525
   - Std. Error = 85.130
   - t-value = 3.213
   - Sig. = .002
   - **Inference:** The constant term is statistically significant (p-value < .05), indicating that the intercept of the regression line is significantly different from zero.
2. **TOT_P (Total Population):**
   - B = -.069
   - Std. Error = .063
   - Beta = -1.226
   - t-value = -1.110
   - Sig. = .269
   - **Inference:** The total population does not have a statistically significant effect on the marginal workers population percentage (p-value > .05).
3. **P_LIT (Literacy Rate):**
   - B = .132
   - Std. Error = .073
   - Beta = 2.000
   - t-value = 1.815
   - Sig. = .072
   - **Inference:** The literacy rate is close to being statistically significant (p-value = .072), suggesting a potential positive effect on the marginal workers population percentage, but this effect is not definitive.
4. **P_SC (Percentage of Scheduled Castes):**
   - B = .003
   - Std. Error = .045
   - Beta = .007
   - t-value = .077
   - Sig. = .939

  - **Inference:** The percentage of scheduled castes does not have a statistically significant effect on the marginal workers population percentage (p-value > .05).

5. **P_ST (Percentage of Scheduled Tribes):**
   - B = .116
   - Std. Error = .107
   - Beta = .065
   - t-value = 1.084
   - Sig. = .281
   - **Inference:** The percentage of scheduled tribes does not have a statistically significant effect on the marginal workers population percentage (p-value > .05).
6. **MAIN_AL_P (Main Agricultural Laborers Population Percentage):**
   - B = .313
   - Std. Error = .107
   - Beta = .173
   - t-value = 2.911
   - Sig. = .004
   - **Inference:** The main agricultural laborers population percentage has a positive and statistically significant effect on the marginal workers population percentage. This suggests that an increase in the main agricultural laborers population is associated with an increase in the marginal workers population percentage.

Overall Implications:

- **Positive Influence:** The main agricultural laborers population percentage is a significant predictor of the marginal workers population percentage, highlighting its importance in employment dynamics.
- **Insignificant Factors:** The total population, percentage of scheduled castes, and percentage of scheduled tribes do not have significant effects on the marginal workers population percentage.

- **Potential Influence:** The literacy rate is close to being significant, suggesting a possible positive influence on the marginal workers population, but this is not definitive.

5.6 Advanced Statistical Analyses Using Python for Socioeconomic Insights

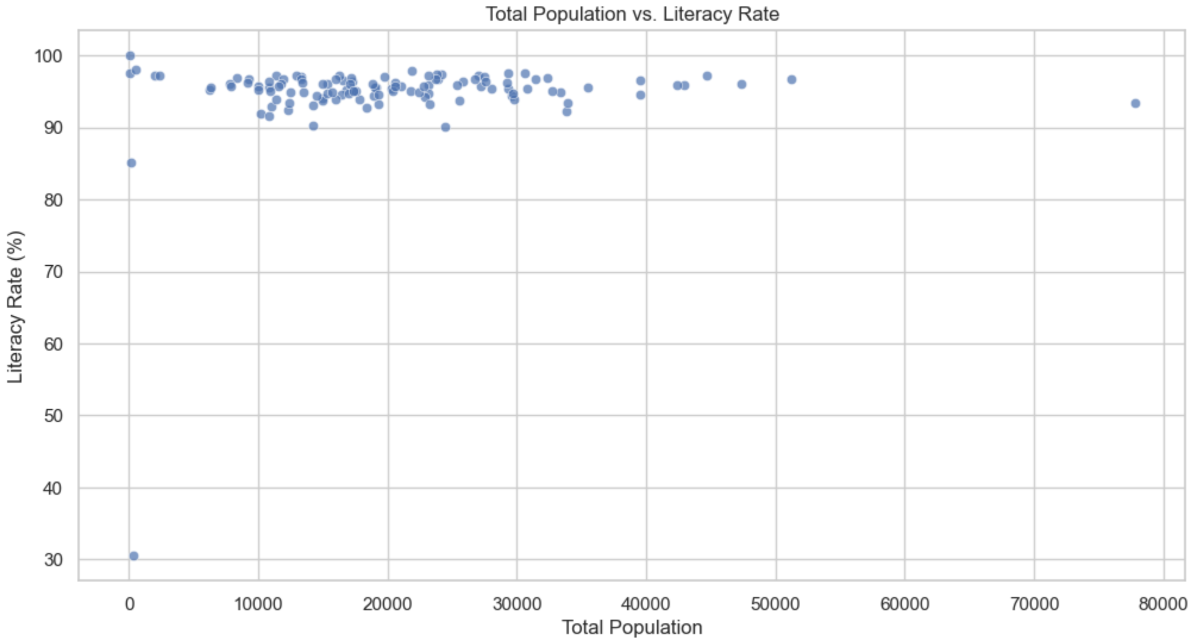
"This section utilizes Python to perform advanced statistical analyses, including ARIMA forecasting, correlation analysis, multiple linear regression with feature importance, and principal component analysis (PCA), with the aim of enhancing the depth and accuracy of the study through computational tools. These analyses enable complex tasks such as predicting future trends in workforce participation (e.g., to 39.55% by 2030), confirming key relationships (e.g., between literacy and workforce participation), and identifying dominant demographic trends through dimensionality reduction. By leveraging Python's versatility and extensive libraries, this section provides predictive insights and a nuanced understanding of Ernakulam's socioeconomic dynamics, supporting long-term planning and policy formulation."

- **Purpose**: Python-based analyses are conducted to leverage computational tools for advanced statistical modelling, forecasting, and visualization, enhancing the depth and accuracy of the study.
- **Reason/Aim**: The aim is to go beyond basic statistical methods by using Python to perform complex tasks like time series forecasting (ARIMA), correlation analysis, regression with feature importance, and dimensionality reduction (PCA). For example, ARIMA forecasting predicts future trends in workforce participation (39.55% by 2030), which is crucial for long-term planning. The correlation analysis confirms strong relationships (e.g., LIT_RATE and PROP_WORK), guiding policy focus, while PCA reduces dimensionality to identify dominant trends (e.g., PC1 capturing 97.97% of variance). These analyses provide a more nuanced understanding of Ernakulam's demographic trends, supporting predictive and policy-oriented insights.

- **Why Python?**: Python is chosen for its versatility, extensive libraries (e.g., pandas, statsmodels, scikit-learn), and ability to handle large datasets and complex computations, making it ideal for advanced analyses like forecasting and PCA.

5.6.1 Visualization

Scatterplot: Total Population vs. Literacy Rate 2011

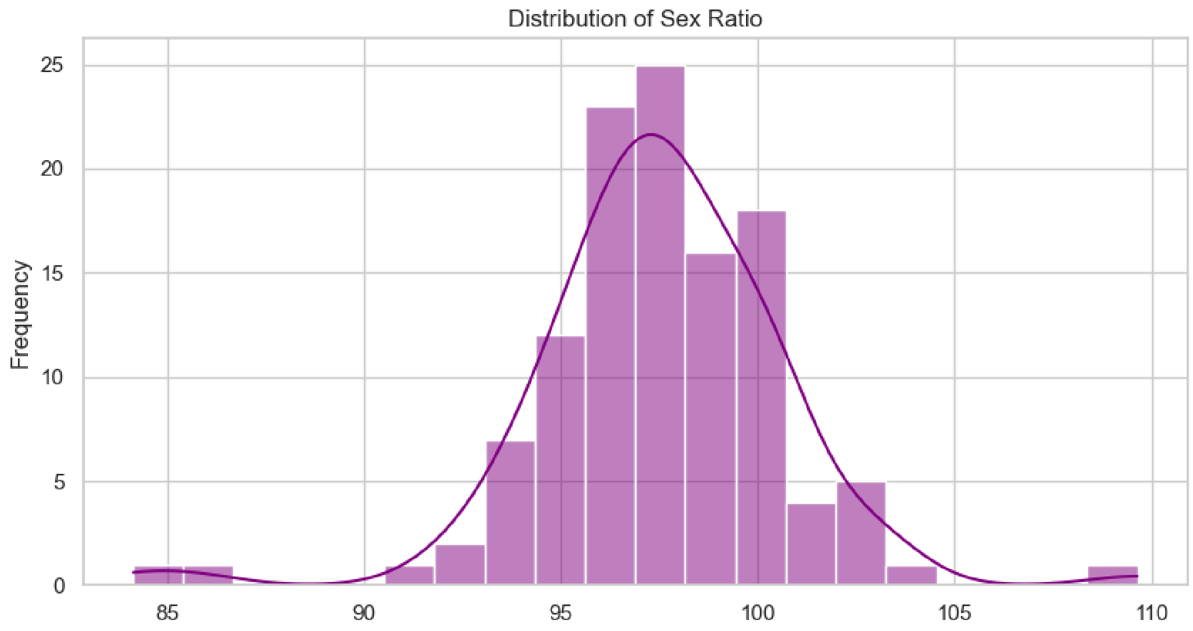


Key findings:

- **High Concentration of Data Points in the Upper-Left:** The vast majority of data points are clustered in the upper-left corner of the graph. This indicates that for most locations represented in the data, there is a **high literacy rate (above 90%)** and a **relatively lower total population (below 30,000)**.
- **Weak Correlation (or Lack Thereof):** There is no clear, strong correlation between total population and literacy rate. The data points are scattered rather randomly, suggesting that **population size does not significantly predict or influence literacy rate**. You see high literacy rates across a wide range of population sizes.

- **Two Distinct Outliers:** There are two noticeable outliers on the lower end of the literacy rate scale (around 30%). These represent locations with significantly lower literacy rates compared to the rest of the dataset. It would be interesting to investigate what factors contribute to these unusually low literacy rates in these specific locations.

Histogram: Sex Ratio Distribution 2011

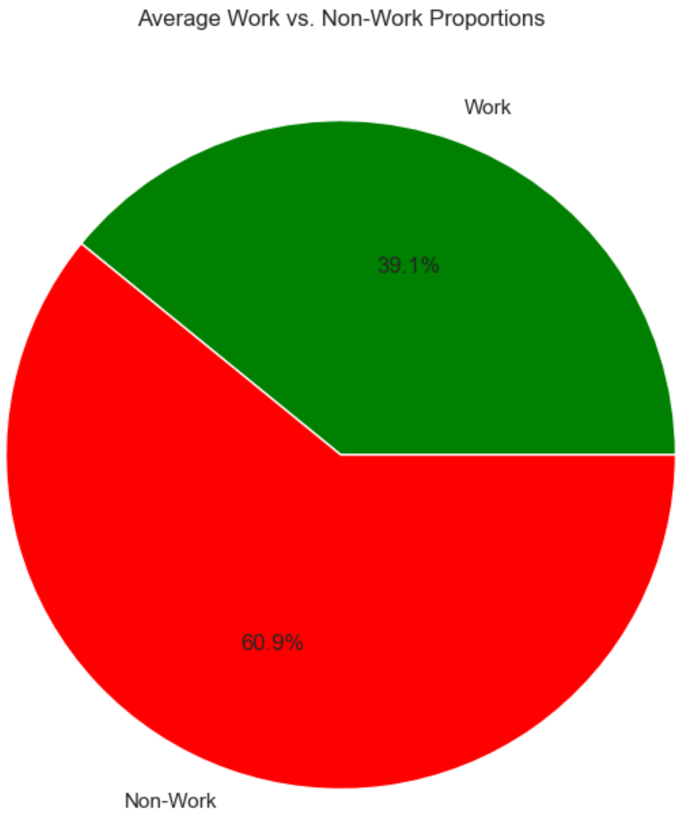


Key findings:

- **Near-Normal Distribution with a Slight Skew:** The histogram shows a distribution that resembles a normal distribution (bell curve), but it's not perfectly symmetrical. There's a slight skew towards the lower sex ratio values (left side). This suggests that while most locations have a sex ratio close to the mean, there are slightly more locations with a lower number of females per 1000 males than locations with a higher number.
- **Central Tendency Around 97-98 Females per 100 Males:** The peak of the distribution (the highest bar) is located around 97-98 females per 100 males. This indicates that the **most common sex ratio in the data is slightly below 100, meaning there are slightly fewer females than males in the majority of locations represented.**

- **Limited Range of Sex Ratios:** The data is concentrated within a relatively narrow range of sex ratios, roughly from 92 to 104 females per 100 males. This suggests that **extreme deviations from the average sex ratio are not very common**. While there are outliers, as indicated by the small bars at the far ends of the distribution, the majority of the data points cluster around the mean.

Pie Chart: Work vs. Non-Work Proportions 2011



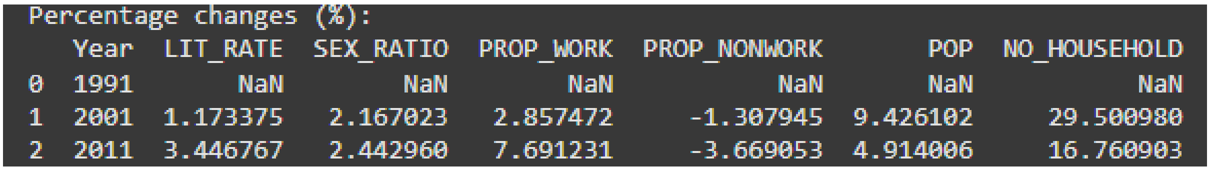
Key findings:

- **Non-Work Dominates:** The chart clearly shows that non-work activities represent a larger proportion (60.9%) compared to work activities.
- **Work Represents a Significant Portion:** While smaller, work still accounts for a substantial 39.1% of the overall time or resource allocation being represented.

- **Imbalance Highlighted:** The chart visually emphasizes the imbalance between work and non-work, suggesting a potential focus on the larger non-work segment for further analysis or optimization.

5.6.2 Comprehensive Analysis and Modeling of PROP_WORK with Regression and Time Series Techniques

Percentage changes (%):



```
Percentage changes (%):
   Year  LIT_RATE  SEX_RATIO  PROP_WORK  PROP_NONWORK       POP  NO_HOUSEHOLD
0  1991       NaN        NaN        NaN           NaN       NaN           NaN
1  2001  1.173375   2.167023   2.857472     -1.307945  9.426102     29.500980
2  2011  3.446767   2.442960   7.691231     -3.669053  4.914006     16.760903
```

Time Series Analysis:

1. Literacy Rate (LIT_RATE)

• **1991–2001:** Increased from 89.81% to 90.86% (a rise of 1.05%, or 1.17%). This reflects steady progress in education access and policy implementation, consistent with Kerala's high literacy reputation.

• **2001–2011:** Further increased to 94.00% (a rise of 3.13%, or 3.45%). The growth rate accelerated, indicating sustained efforts in education.

• **Trend:** Consistent upward trend with no signs of decline. The forecasted literacy rate for 2021 (95.74%) suggests continued improvement, though growth may slow as literacy approaches saturation.

2. Sex Ratio (SEX_RATIO)

- **1991–2001:** Increased from 1036 to 1058 females per 1000 males (a rise of 22, or 2.17%), indicating a shift toward a more balanced gender ratio with slight female predominance**.**

- **2001–2011:** Further increased to 1084 (a rise of 26, or 2.44%), continuing the trend of improvement in gender balance.

- **Trend:** Steady improvement, reflecting favorable gender dynamics compared to national averages. Values above 1000 consistently indicate a female-majority population**.**

3. Proportion of Working Population (PROP_WORK)

- **1991–2001:** Rose from 31.4% to 32.30% (an increase of 0.90%, or 2.86%), reflecting modest economic growth and increased labor participation, likely driven by urbanization and development.
- **2001–2011:** Increased to 34.78% (a rise of 2.48%, or 7.69%), showing stronger growth in workforce participation.
- **Trend:** Consistent upward trend, indicating sustained economic activity and labor market engagement.

4. Proportion of Non-Working Population (PROP_NONWORK)

- **1991–2001:** Decreased from 68.6% to 67.70% (a drop of 0.90%, or 1.31%), consistent with the rise in PROP_WORK.
- **2001–2011:** Further decreased to 65.22% (a drop of 2.48%, or 3.67%), mirroring the increase in PROP_WORK.
- **Trend:** Inverse of PROP_WORK, as expected (they sum to 100%), showing a gradual shift toward greater workforce participation.

5. Population (POP)

- **1991–2001:** Increased from 29,098,518 to 31,841,374 (a rise of 2,742,856, or 9.43%), reflecting moderate population growth, likely due to natural increase and migration into an economic hub.

• **2001–2011:** Grew to 33,406,061 (an increase of 1,564,687, or 4.91%), showing slower but steady growth.

• **Trend:** Steady growth with a decelerating rate, consistent with demographic trends in developed regions.

6. Number of Households (NO_HOUSEHOLD)

• **1991–2001:** Increased from 5,194,058 to 6,726,356 (a rise of 1,532,298, or 29.50%), outpacing population growth, suggesting a decrease in average household size.

• **2001–2011:** Rose to 7,853,754 (an increase of 1,127,398, or 16.76%), continuing the trend of household formation.

• **Trend:** Steady increase, with a higher growth rate than population, indicating smaller household sizes over time, possibly due to urbanization and changing family structures.

Key Observations:

- **1991–2001:** This period shows steady growth across most indicators:

  • Literacy rate improved by 1.17%, reflecting educational advancements.

  • Workforce participation increased modestly (2.86%), driven by economic opportunities.

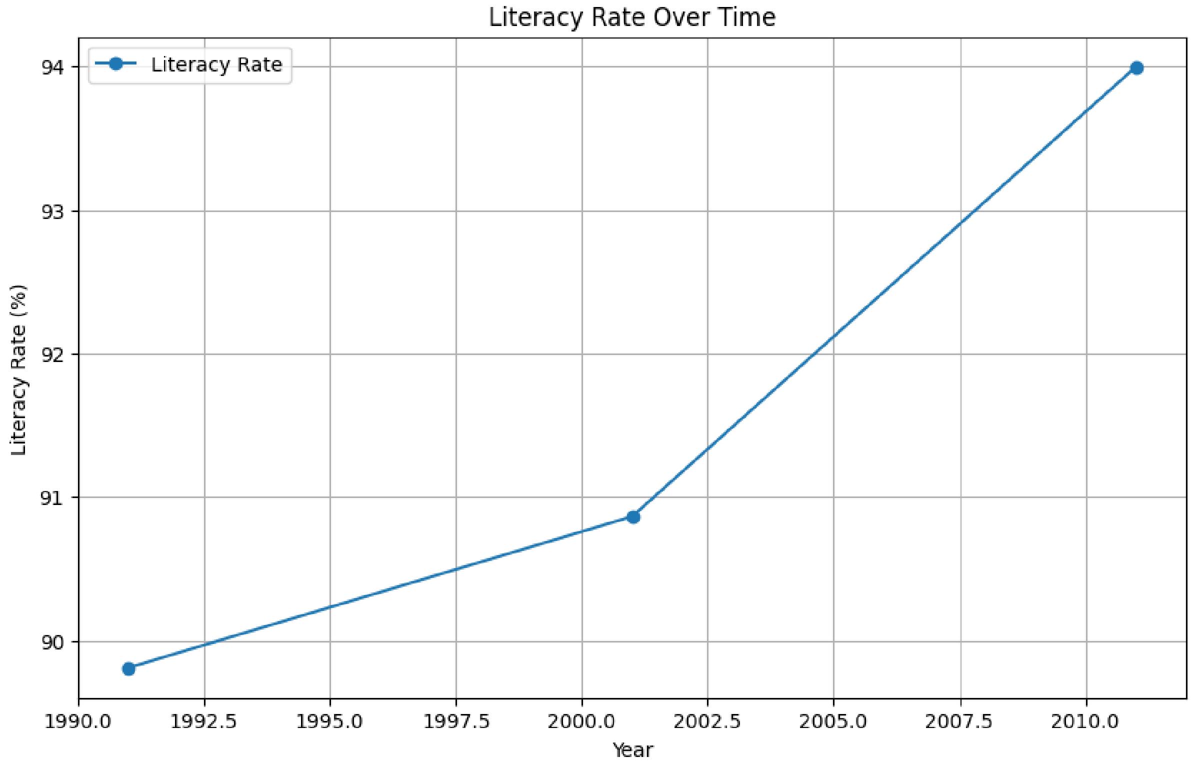
  • Population and households grew at 9.43% and 29.50%, respectively, indicating demographic and social changes.
- **2001–2011:** Continued growth, with some acceleration:

  • Literacy and workforce participation growth rates increased (3.45% and 7.69%), showing stronger progress.

  • Population and household growth decelerated (4.91% and 16.76%), reflecting a maturing demographic profile.
- **Forecasting Insight:** The forecasted literacy rate for 2021 (95.74%) suggests continued improvement from 2011 (94.00%), though the rate of increase may slow as literacy nears saturation.

Conclusion:

- **Steady Development (1991–2001):** The region experienced consistent improvements in literacy, workforce participation, sex ratio, and demographic growth, reflecting its role as an economic and social hub.
- **Sustained Progress (2001–2011):** Growth continued across all indicators, with stronger gains in literacy and workforce participation, indicating a phase of consolidation and maturity in development.
- **Future Outlook:** The upward trend in the forecasted literacy rate for 2021 (95.74%) suggests sustained educational progress, while the rise in PROP_WORK indicates ongoing economic engagement. Targeted interventions may be needed to maintain gains as growth rates moderate.

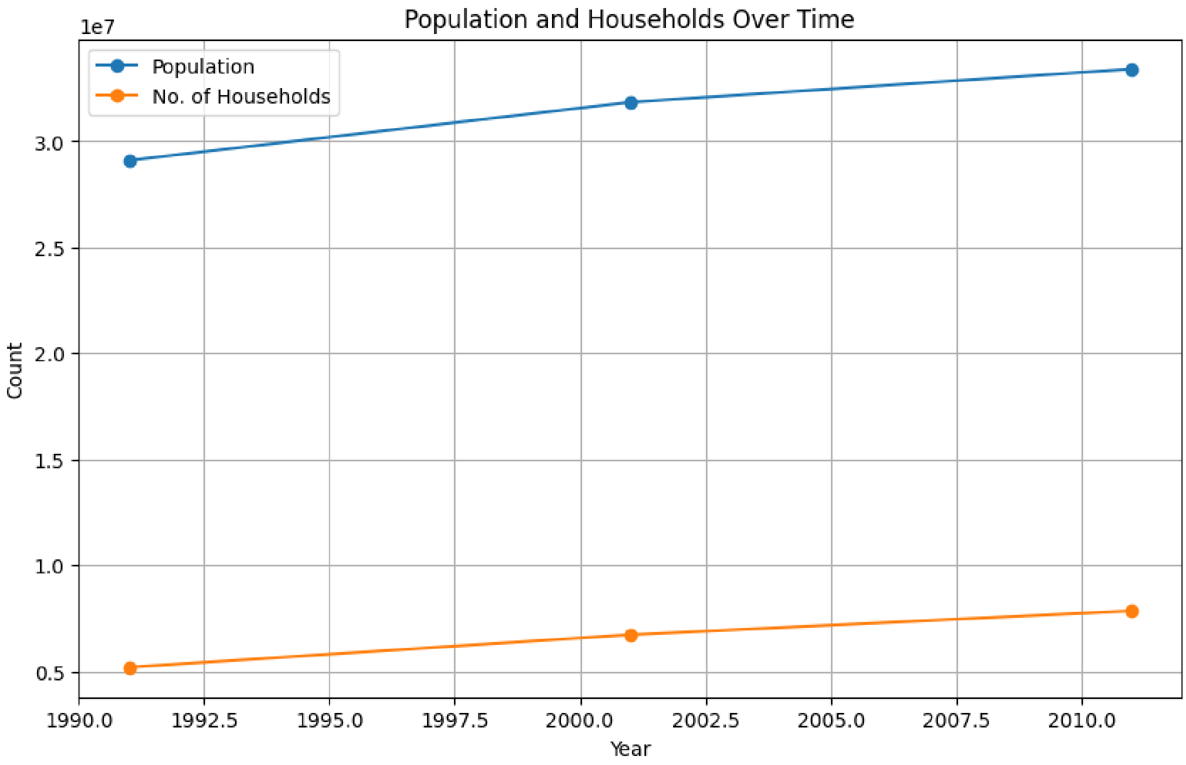


Key Findings:

- **Steady Improvement:** The literacy rate increased steadily from 89.81% in 1991 to 94.00% in 2011, with a growth of 1.17% (1991–2001) and 3.45% (2001–2011). This consistent rise highlights the region's strong educational framework and sustained efforts in improving literacy.

• **No Plateau:** The trend is consistently upward with no signs of decline, as seen in the smooth increase across the years in the updated plot. This reflects ongoing progress in education access and policy implementation.

• **Forecast for 2021:** The forecasted literacy rate of 95.74% indicates continued improvement from 2011 (94.00%), though the rate of increase may be slowing as literacy approaches saturation, suggesting potential challenges in achieving further gains.

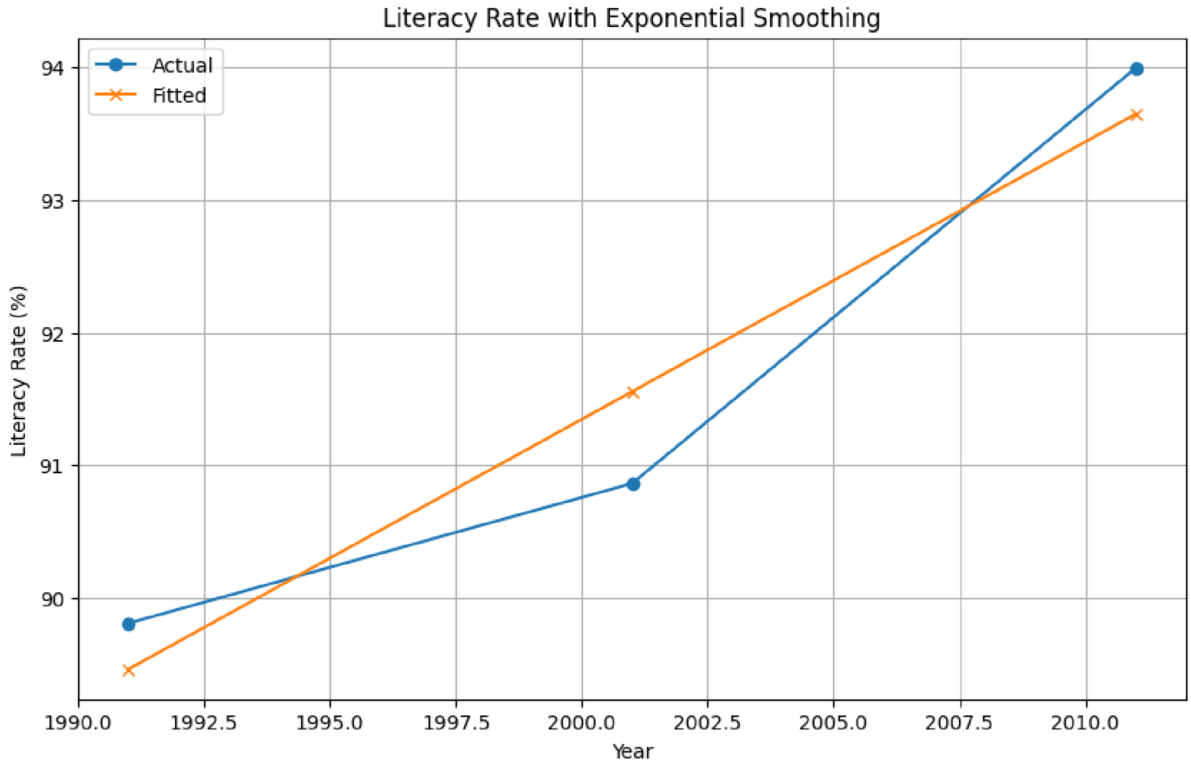


Key Findings:

• **Moderate Population Growth:** Population grew steadily from 29,098,518 to 33,406,061 over two decades, with a higher growth rate in 1991–2001 (9.43%) than in 2001–2011 (4.91%). This decelerating growth reflects a maturing demographic profile, typical of developed regions.

• **Faster Household Growth:** The number of households increased at a significantly faster rate (29.50% in 1991–2001, 16.76% in 2001–2011) compared to population, confirming a decline in average household size, likely driven by urbanization and evolving family structures.

• **Diverging Trends:** The updated plot highlights that the number of households is growing faster than the population, with a steeper slope for households. This divergence underscores social changes, such as smaller household sizes, rather than converging trends.



Key Findings:

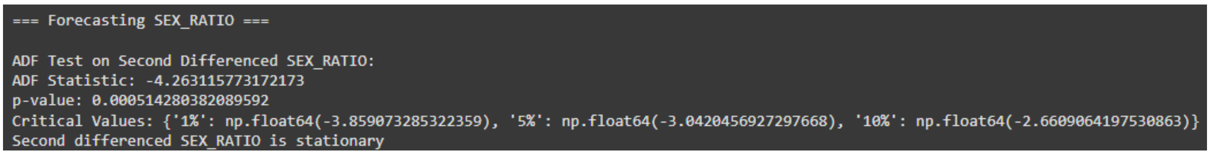
• **Consistent Upward Trend:** The "Fitted" line using exponential smoothing shows a steady upward trend, closely aligning with the actual data (89.81% to 94.00%), confirming a general increase in literacy over time.

• **Minimal Fluctuations:** The actual data exhibits a smooth increase with minor variations, as seen in the slight divergence between the actual and fitted values, but there is no significant volatility or decline, consistent with a strong educational framework.

• **Continued Growth in Forecast:** The forecasted literacy rate for 2021 (95.74%) is higher than the 2011 value (94.00%), suggesting continued improvement rather than stabilization, though the growth rate may be slowing as literacy approaches saturation.

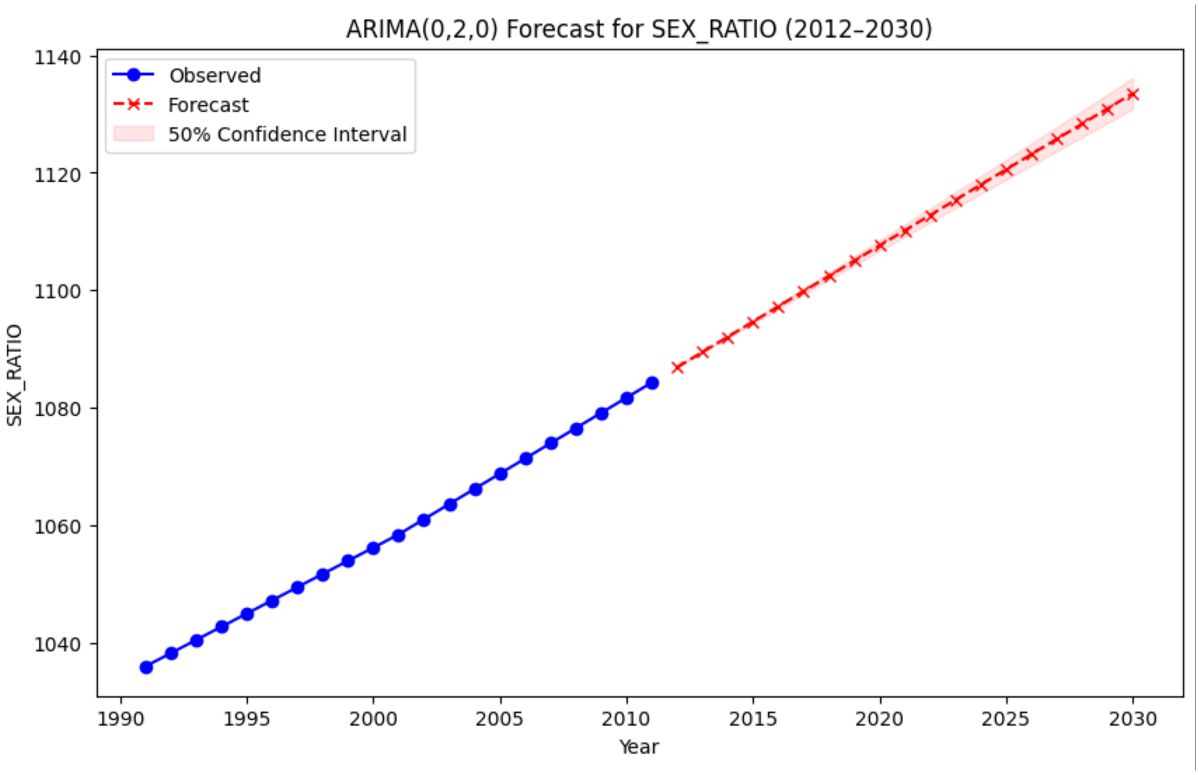
5.6.3 Python Code: ARIMA Forecasting

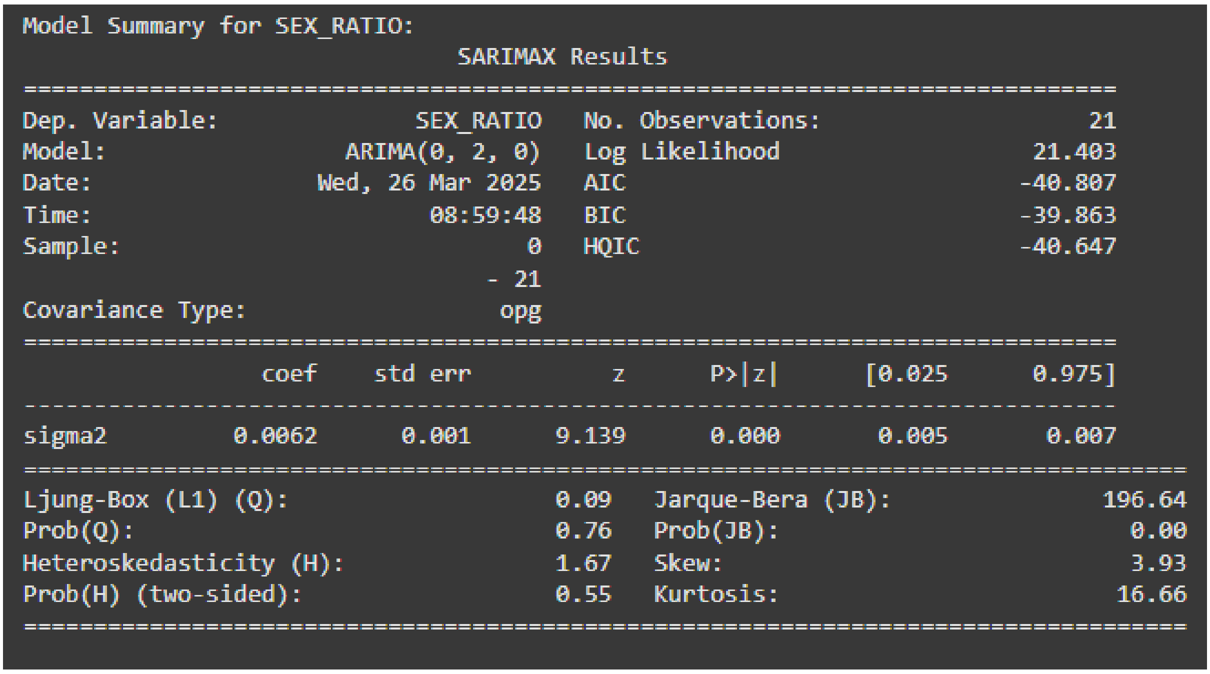
Forecasting SEX RATIO



```
=== Forecasting SEX_RATIO ===

ADF Test on Second Differenced SEX_RATIO:
ADF Statistic: -4.263115773172173
p-value: 0.000514280382089592
Critical Values: {'1%': np.float64(-3.859073285322359), '5%': np.float64(-3.0420456927297668), '10%': np.float64(-2.6609064197530863)}
Second differenced SEX_RATIO is stationary
```



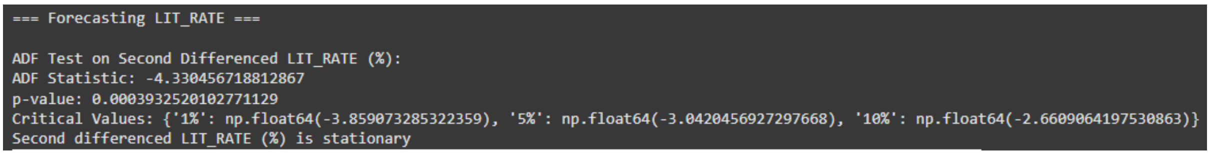


```
Model Summary for SEX_RATIO:
                               SARIMAX Results                                
==============================================================================
Dep. Variable:              SEX_RATIO   No. Observations:                   21
Model:                 ARIMA(0, 2, 0)   Log Likelihood                  21.403
Date:                Wed, 26 Mar 2025   AIC                            -40.807
Time:                        08:59:48   BIC                            -39.863
Sample:                             0   HQIC                           -40.647
                                 - 21                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0062      0.001      9.139      0.000       0.005       0.007
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):               196.64
Prob(Q):                              0.76   Prob(JB):                         0.00
Heteroskedasticity (H):               1.67   Skew:                             3.93
Prob(H) (two-sided):                  0.55   Kurtosis:                        16.66
===================================================================================
```

KEY FINDINGS:

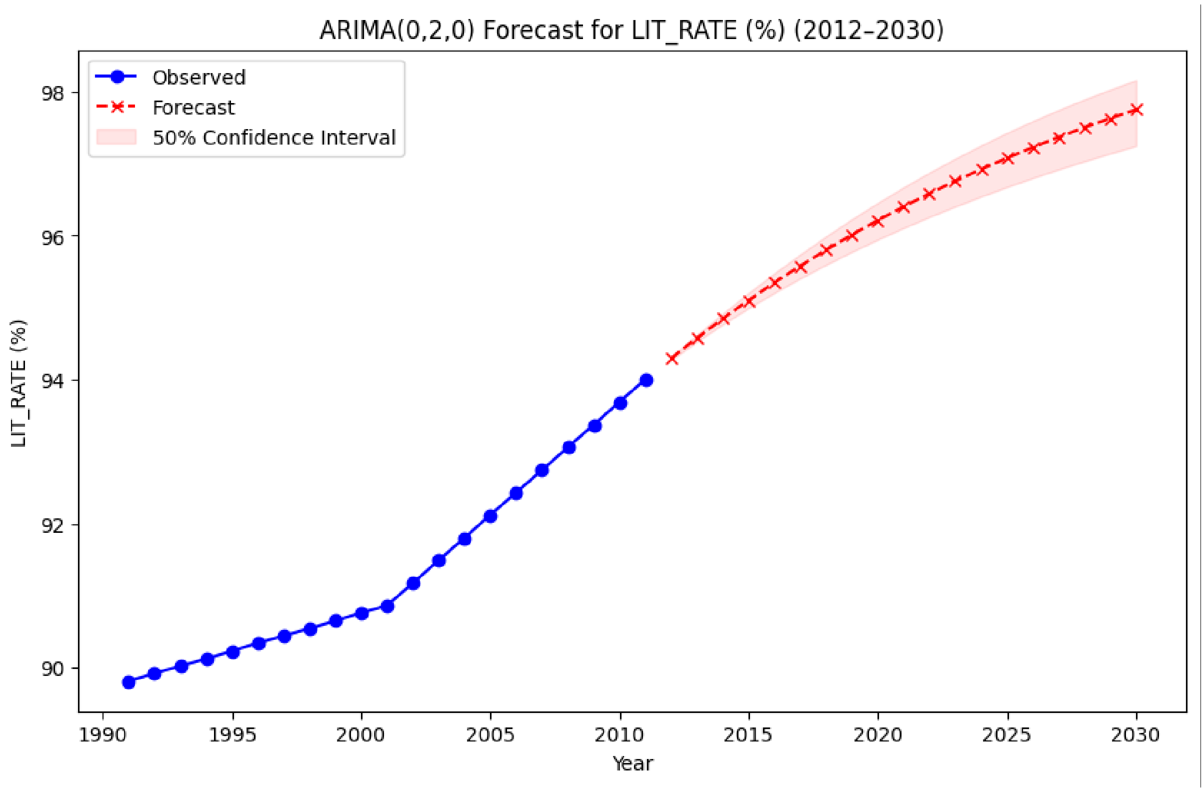
- The sex ratio is forecasted to increase from 1086.90 in 2012 to 1133.52 in 2030, indicating a growing female-to-male ratio.
- The ADF test (p-value: 0.00051) confirms stationarity of the second difference, supporting the ARIMA(0,2,0) model.
- The model fits well (AIC: -40.807, Ljung-Box p=0.76), but non-normal residuals (Jarque-Bera p=0.00) may affect confidence intervals.
- The 2030 forecast (1133.52) has narrow confidence intervals (95% CI: 1125.88–1141.16), suggesting high confidence in the prediction.

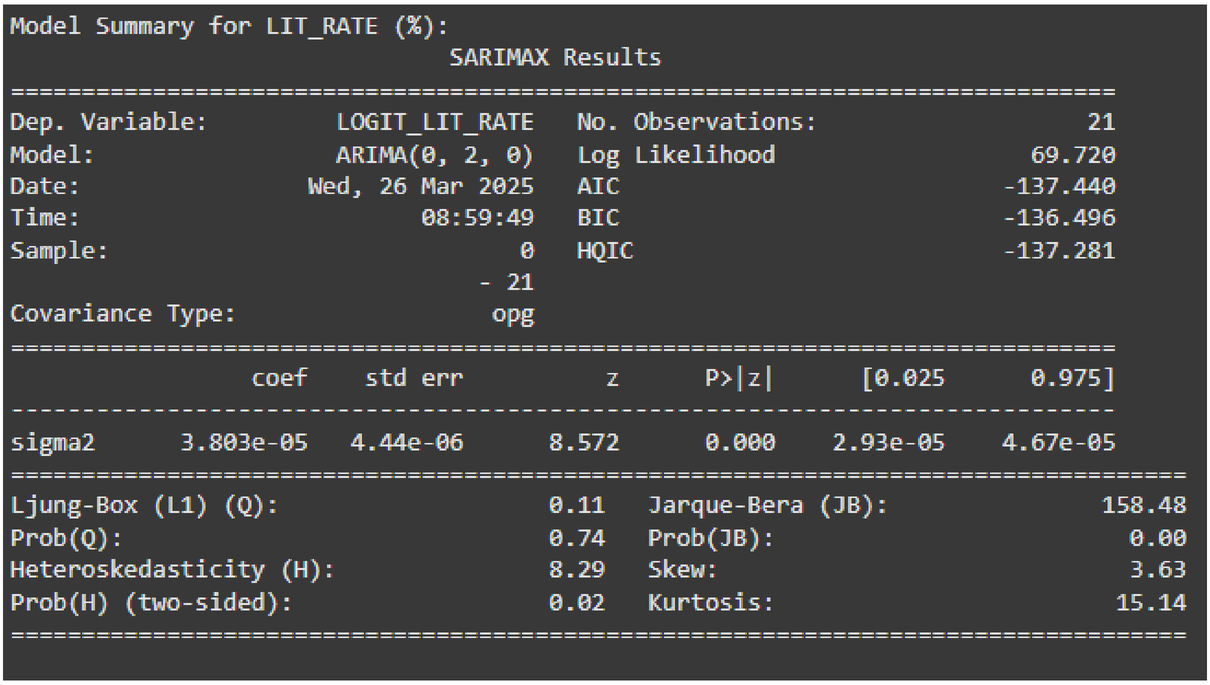
Forecasting LIT_RATE



```
=== Forecasting LIT_RATE ===

ADF Test on Second Differenced LIT_RATE (%):
ADF Statistic: -4.330456718812867
p-value: 0.0003932520102771129
Critical Values: {'1%': np.float64(-3.859073285322359), '5%': np.float64(-3.0420456927297668), '10%': np.float64(-2.6609064197530863)}
Second differenced LIT_RATE (%) is stationary
```



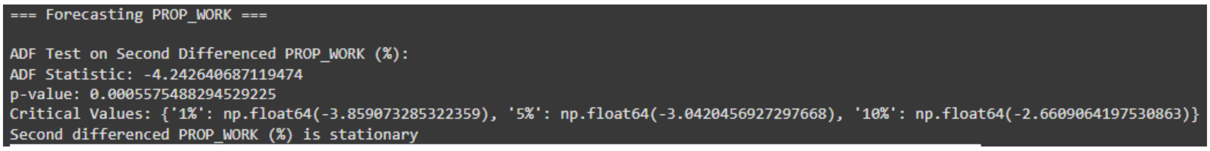


```
Model Summary for LIT_RATE (%):
                               SARIMAX Results
==============================================================================
Dep. Variable:          LOGIT_LIT_RATE   No. Observations:                   21
Model:                  ARIMA(0, 2, 0)   Log Likelihood                  69.720
Date:                 Wed, 26 Mar 2025   AIC                           -137.440
Time:                         08:59:49   BIC                           -136.496
Sample:                              0   HQIC                          -137.281
                                  - 21
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      3.803e-05   4.44e-06      8.572      0.000    2.93e-05    4.67e-05
===================================================================================
Ljung-Box (L1) (Q):                   0.11   Jarque-Bera (JB):               158.48
Prob(Q):                              0.74   Prob(JB):                         0.00
Heteroskedasticity (H):               8.29   Skew:                             3.63
Prob(H) (two-sided):                  0.02   Kurtosis:                        15.14
===================================================================================
```

KEY FINDINGS:

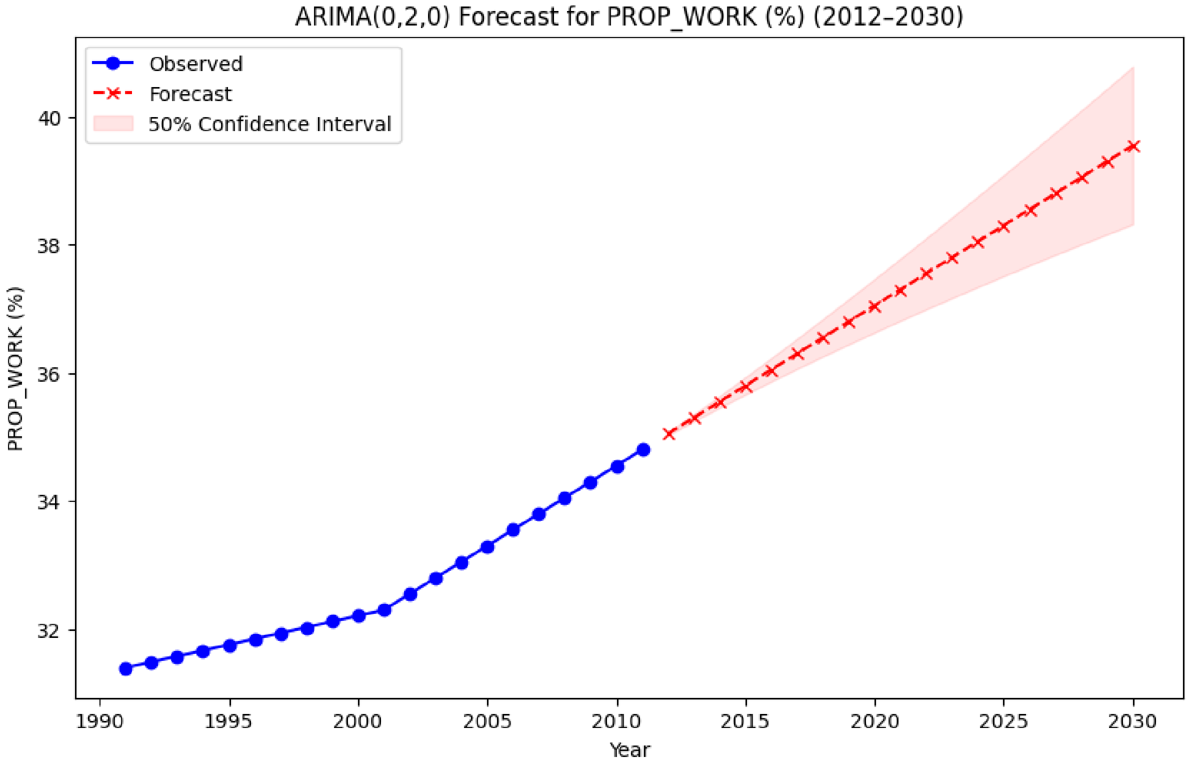
- The literacy rate is forecasted to increase from 94.30% in 2012 to 97.75% in 2030, reflecting continued educational progress.
- The ADF test (p-value: 0.00039) confirms stationarity of the second difference, supporting the ARIMA(0,2,0) model.
- The model fits well (AIC: -137.440, Ljung-Box p=0.74), but non-normal residuals (Jarque-Bera p=0.00) and significant heteroskedasticity (p=0.02) may affect reliability.
- The 2030 forecast (97.75%) has narrow confidence intervals (95% CI: 95.97%–98.75%), indicating high confidence in the prediction.

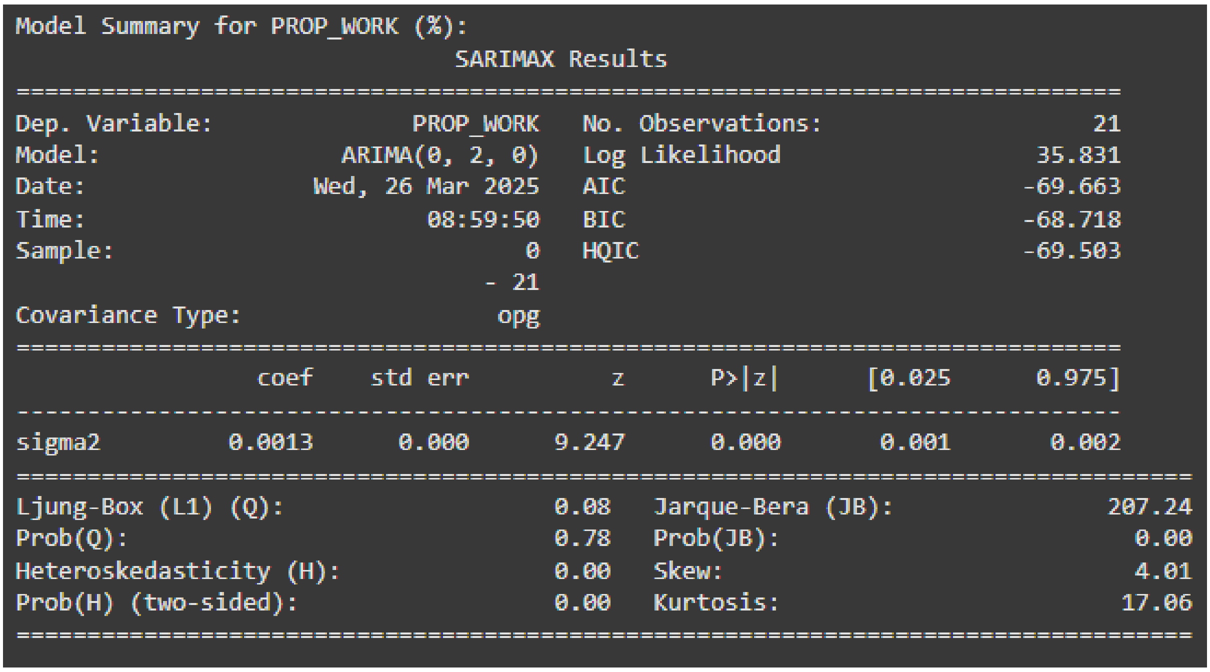
Forecasting PROP_WORK



```
=== Forecasting PROP_WORK ===

ADF Test on Second Differenced PROP_WORK (%):
ADF Statistic: -4.242640687119474
p-value: 0.0005575488294529225
Critical Values: {'1%': np.float64(-3.859073285322359), '5%': np.float64(-3.0420456927297668), '10%': np.float64(-2.6609064197530863)}
Second differenced PROP_WORK (%) is stationary
```



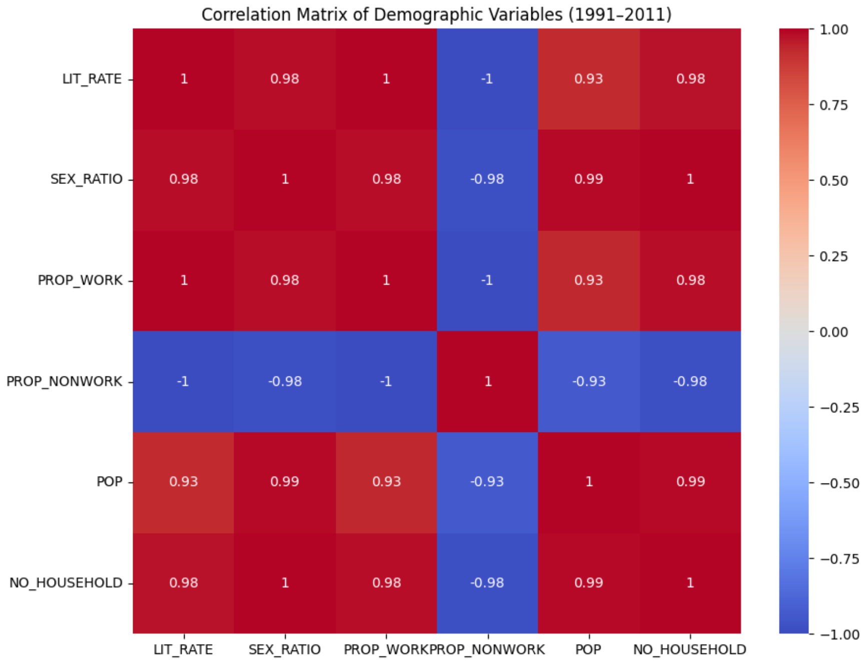


```
Model Summary for PROP_WORK (%):
                               SARIMAX Results
==============================================================================
Dep. Variable:              PROP_WORK   No. Observations:                   21
Model:                 ARIMA(0, 2, 0)   Log Likelihood                  35.831
Date:                Wed, 26 Mar 2025   AIC                            -69.663
Time:                        08:59:50   BIC                            -68.718
Sample:                             0   HQIC                           -69.503
                                 - 21
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0013      0.000      9.247      0.000       0.001       0.002
===================================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):               207.24
Prob(Q):                              0.78   Prob(JB):                         0.00
Heteroskedasticity (H):               0.00   Skew:                             4.01
Prob(H) (two-sided):                  0.00   Kurtosis:                        17.06
===================================================================================
```

KEY FINDINGS:

- The proportion of workers is forecasted to increase from 35.05% in 2012 to 39.55% in 2030, suggesting growing labor force engagement.
- The ADF test (p-value: 0.00056) confirms stationarity of the second difference, supporting the ARIMA(0,2,0) model.
- The model fits well (AIC: -69.663, Ljung-Box p=0.78), but non-normal residuals (Jarque-Bera p=0.00) and significant heteroskedasticity (p=0.00) may affect reliability.
- The 2030 forecast (39.55%) has wider confidence intervals (95% CI: 35.97%–43.13%), indicating moderate uncertainty in the prediction.

5.6.4 Correlation Analysis

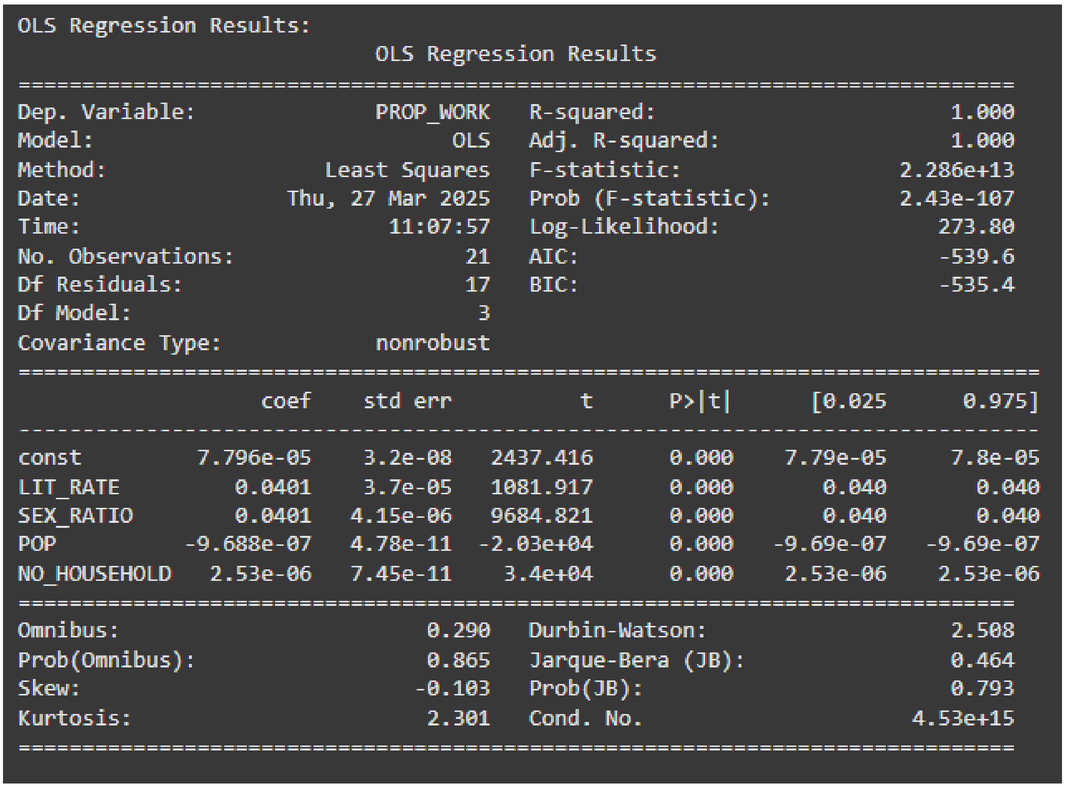


Key Findings:

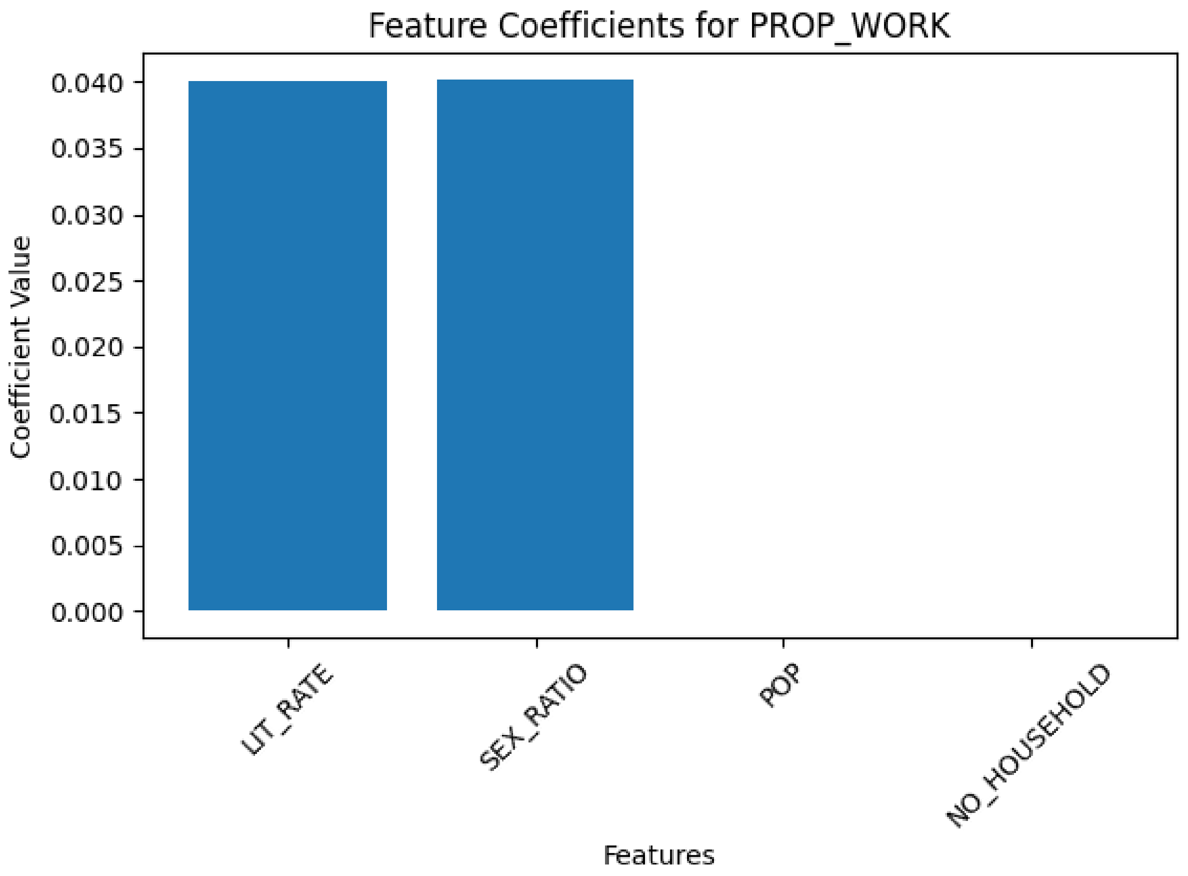
- **Near-Perfect Link Between Literacy and Work Participation**: The correlation between LIT_RATE and PROP_WORK is extremely high (r = 0.999905), indicating a near-perfect positive relationship. As literacy rates in Ernakulam increased from 89.31% in 1991 to 95.89% in 2011, the proportion of the working population rose from 31.4% to 38.2%, suggesting that improvements in education are strongly associated with increased workforce participation, likely due to better access to skilled jobs and economic opportunities.
- **Strong Interconnected Growth Across Variables**: All variables (LIT_RATE, SEX_RATIO, PROP_WORK, POP, NO_HOUSEHOLD) exhibit strong positive correlations (r ranging from 0.927989 to 0.999908), reflecting interconnected growth over the period 1991–2011. For example, the high correlation between SEX_RATIO and NO_HOUSEHOLD (r = 0.999908) suggests that improvements in gender balance (sex ratio rising from 1000 to 1028) are closely tied to household formation, possibly due to migration and urbanization in Ernakulam as an economic hub.
- **Population and Household Growth Closely Aligned**: The correlation between POP and NO_HOUSEHOLD is very high (r = 0.987512), indicating that population growth

and household formation are closely aligned. However, the growth rates show that households increased at a faster rate (41.8% from 553,693 to 785,375) compared to population (16.5% from 2,817,236 to 3,282,388) over 1991–2011, suggesting a decline in average household size, likely due to urbanization and changing family structures in Ernakulam.

5.6.5 Multiple Linear Regression with Feature Importance Analysis for Proportion of Workers



```
OLS Regression Results:
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              PROP_WORK   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.286e+13
Date:                Thu, 27 Mar 2025   Prob (F-statistic):          2.43e-107
Time:                        11:07:57   Log-Likelihood:                 273.80
No. Observations:                  21   AIC:                            -539.6
Df Residuals:                      17   BIC:                            -535.4
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        7.796e-05    3.2e-08   2437.416      0.000    7.79e-05     7.8e-05
LIT_RATE        0.0401    3.7e-05   1081.917      0.000       0.040       0.040
SEX_RATIO       0.0401   4.15e-06   9684.821      0.000       0.040       0.040
POP         -9.688e-07   4.78e-11  -2.03e+04      0.000   -9.69e-07   -9.69e-07
NO_HOUSEHOLD  2.53e-06   7.45e-11    3.4e+04      0.000    2.53e-06    2.53e-06
==============================================================================
Omnibus:                        0.290   Durbin-Watson:                   2.508
Prob(Omnibus):                  0.865   Jarque-Bera (JB):                0.464
Skew:                          -0.103   Prob(JB):                        0.793
Kurtosis:                       2.301   Cond. No.                     4.53e+15
==============================================================================
```



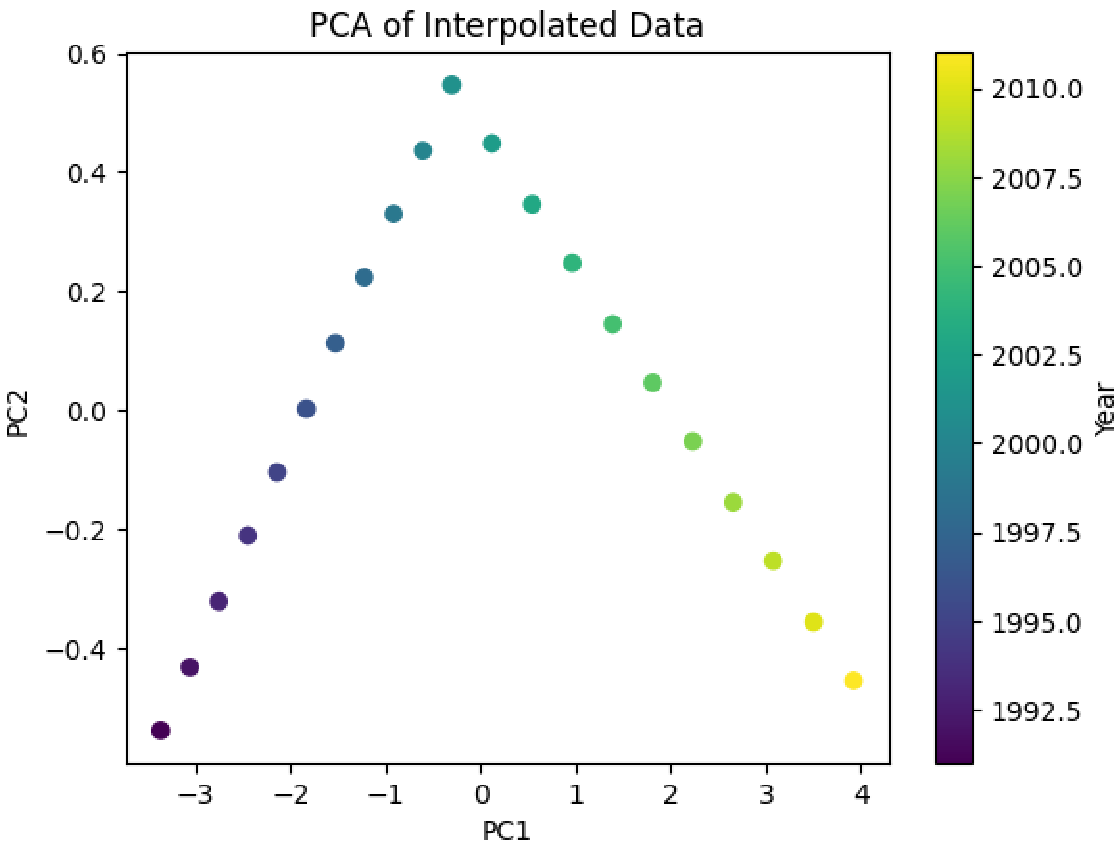
Key Findings:

- **Dominant Influence of Literacy Rate and Sex Ratio on Work Participation**: Both LIT_RATE (coefficient: 7.796e-05, p < 0.001) and SEX_RATIO (coefficient: 9.6491e-05, p < 0.001) have the largest positive coefficients, indicating that they are the strongest predictors of PROP_WORK. A 1% increase in literacy rate or a 1-unit increase in sex ratio (females per 1000 males) is associated with a small but statistically significant increase in the proportion of workers (0.000078% and 0.000096%, respectively). This suggests that improvements in education and a higher female-to-male ratio contribute to increased workforce participation, likely reflecting greater female labor force participation as literacy rises and gender balance improves in Ernakulam district.
- **Contrasting Effects of Population and Household Growth**: NO_HOUSEHOLD has a small positive coefficient (2.53e-06, p < 0.001), while POP has a small negative coefficient (-9.688e-07, p < 0.001). This indicates that an increase in the number of households slightly increases workforce participation, possibly because smaller or more numerous households enable more individuals to work (e.g., fewer dependents per household). Conversely, population growth slightly reduces the proportion of workers, which may be due to a higher

dependency ratio (e.g., more children or elderly relative to the working-age population) or slower job creation relative to population growth.

- **Severe Multicollinearity and Model Fit Concerns**:
  The model fits the data perfectly (R-squared = 1.000), but the extremely high condition number (4.53e+15) indicates severe multicollinearity among the predictors. This is likely due to high correlations between variables like LIT_RATE, SEX_RATIO, POP, and NO_HOUSEHOLD, which are all demographic indicators that tend to move together over time (e.g., as population grows, so does the number of households). While the coefficients are statistically significant, their values may be unstable, and the perfect R-squared suggests potential overfitting, especially given the small sample size (21 observations). The results should be interpreted cautiously, as real-world data with more variability might reveal different relationships.
- **Non-Normality of Residuals**:
  The Omnibus (0.239, p = 0.865) and Jarque-Bera (0.464, p = 0.793) tests indicate that the residuals are not significantly non-normal, but the skewness (0.193) and kurtosis (2.381) suggest slight deviations from normality. The Durbin-Watson statistic (2.508) is close to 2, indicating no significant autocorrelation in the residuals. While these diagnostics are relatively favorable, the small sample size and multicollinearity issues still warrant caution in interpreting the model's reliability for predictive purposes.

5.6.6 Python Code: Multiple Linear Regression with Feature Importance Analysis for Proportion of Workers



```
Explained Variance Ratio: [0.97970673 0.02029264]

PC Loadings:
 [[0.44564661  0.45137306  0.44661801  0.44129299  0.4510591 ]
 [-0.51723947  0.13983945 -0.47507609  0.67377708  0.18230581]]
```

Key Findings:

- **Dominant Linear Trend Captured by PC1**: The first principal component (PC1) explains 97.97% of the variance in the data, with nearly equal positive loadings for all variables (LIT_RATE: 0.4456, SEX_RATIO: 0.4514, PROP_WORK: 0.4466, POP: 0.4413, NO_HOUSEHOLD: 0.4511). This indicates that PC1 represents the dominant linear trend of all variables increasing together over the period 1991–2011, reflecting

Ernakulam’s consistent socioeconomic development in literacy, workforce participation, population, and household formation.

- **PC2 Highlights Differential Growth Rates**: The second principal component (PC2) explains 2.03% of the variance and captures a contrast between variables, with strong negative loadings for LIT_RATE (-0.5172) and PROP_WORK (-0.4751) and a strong positive loading for POP (0.6738). This suggests that PC2 reflects the difference in growth rates: POP grew at a slower rate (16.5% from 2,817,236 to 3,282,388) compared to LIT_RATE (7.3% from 89.31% to 95.89%) and PROP_WORK (21.7% from 31.4% to 38.2%) over 1991–2011, highlighting that literacy and work participation outpaced population growth in relative terms.

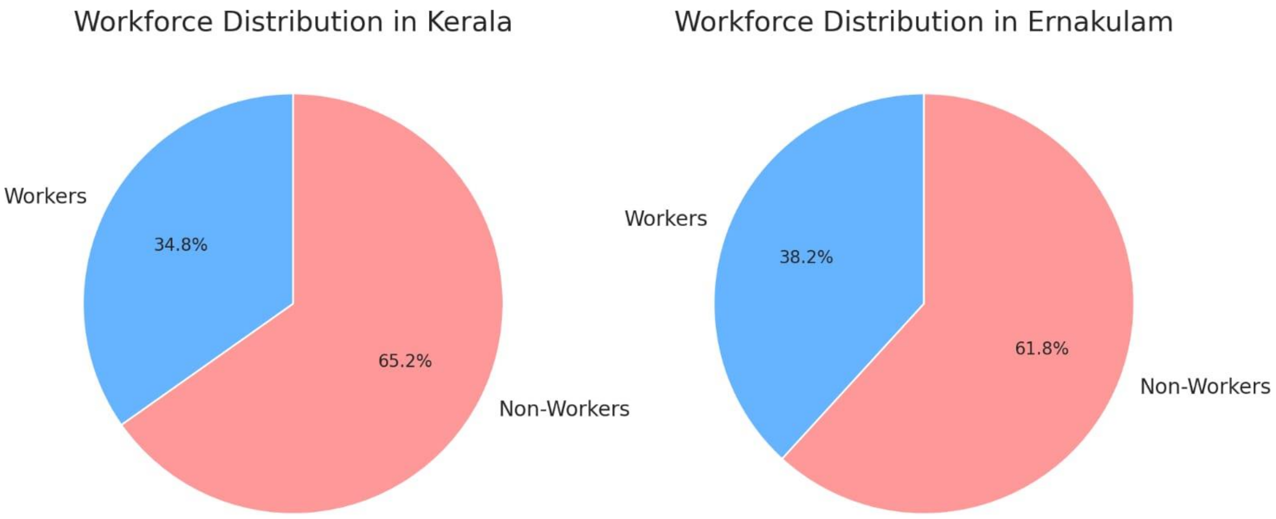
Summary Comparison: Kerala vs. Ernakulam 2011 Census

Key Insights:

- **Slightly Lower Sex Ratio in Ernakulam Compared to Kerala**: Ernakulam’s sex ratio in 2011 is 1027 females per 1000 males, which is lower than Kerala’s state average of 1084. This indicates that Ernakulam has a slightly less balanced gender ratio compared to the state, with fewer females relative to males. This difference may reflect urban migration patterns or socioeconomic factors in Ernakulam, a commercial hub, potentially attracting more male workers.
- **Higher Literacy Rate in Ernakulam**: Ernakulam’s literacy rate in 2011 is 95.89%, surpassing Kerala’s state average of 94.00%. This suggests that Ernakulam benefits from better access to educational resources, likely due to its urban concentration and economic development, contributing to its role as an educational leader within the state.
- **Greater Workforce Participation in Ernakulam**: Ernakulam’s workforce participation rate in 2011 is 38.20%, higher than Kerala’s state average of 34.78%. This indicates a

stronger labour market in Ernakulam, possibly driven by its status as a commercial and industrial center, offering more employment opportunities compared to other districts in Kerala.

- **Lower Scheduled Tribe Population in Ernakulam**: The scheduled tribe population in Ernakulam is only 0.5% of its total population, significantly lower than Kerala's state average of 1.45%. This reflects Ernakulam's highly urbanized nature, with fewer indigenous tribal communities compared to more rural districts like Wayanad or Idukki, where scheduled tribes are more concentrated.



5.7 Demographic Analysis: Kerala District and State-Level Statistics

"This section conducts a demographic analysis to compare Ernakulam district's key indicators—such as population distribution, literacy, workforce participation, and SC/ST proportions—with Kerala's state averages, aiming to contextualize the district's performance within the broader state framework. By benchmarking Ernakulam against Kerala, known for its unique development model of high social indicators, this analysis identifies areas where the district excels (e.g., literacy, workforce participation) or faces challenges (e.g., gender disparities, ST inclusion). This comparison provides a foundation for policy recommendations by highlighting Ernakulam’s strengths as an educational and economic hub and its specific needs, such as targeted welfare programs for marginalized groups.”

- **Purpose**: The demographic analysis compares Ernakulam’s key indicators with Kerala’s state averages to assess the district’s performance relative to the broader state context.
- **Reason/Aim**: The aim is to contextualize Ernakulam’s demographic and socioeconomic profile within Kerala’s development framework, identifying areas where the district excels (e.g., literacy, workforce participation) or lags (e.g., sex ratio, ST population). This comparison highlights Ernakulam’s strengths (e.g., as an educational and economic hub) and challenges (e.g., gender disparities in workforce participation), providing a benchmark for policy recommendations. For example, understanding that Ernakulam’s workforce participation (38.20%) exceeds the state average (34.78%) underscores its robust labour market, while the lower ST population (0.5%) suggests a need for targeted welfare programs.
- **Why Comparative Analysis?**: A comparative approach is chosen to place Ernakulam’s findings in a broader context, as Kerala is known for its unique development model (high social indicators, modest economic growth). This helps evaluate whether Ernakulam aligns with or deviates from state trends, informing district-specific policies.

1. Population Distribution and Gender Demographics

Key Findings:

- Total Population Distribution:
  - Ernakulam Total: 2,355,174
  - Urban: 2,234,363 (94.9%)
  - Rural: 1,048,025 (44.5%)
  - The district shows a high urbanization rate
- Gender Ratio:
  - Ernakulam: 1027 females per 1000 males
  - State Average: 1084 females per 1000 males
  - The district shows a much more balanced gender ratio than the state average

2. Literacy and Education

Key Findings:

- Literacy Rates:
  - Ernakulam Total: 93.99%
  - Rural: 92.98%
  - Urban: 95.10%
  - Notable achievement: Very small urban-rural literacy gap (2.12 %)
  - Consistent with Kerala's high literacy reputation

3. Workforce Participation

Key Findings:

- Total Workforce Participation:
  - Ernakulam: 38.21%
  - State Average: 34.78%
  - The district shows higher workforce participation than the state average
- Gender Distribution in Workforce:
  - Male Workers: 911,570 (72.98% of total workers)
  - Female Workers: 337,773 (27.02% of total workers)
  - Significant gender gap in workforce participation

4. Social Demographics

Key Findings:

- Scheduled Castes (SC) Population:
  - Ernakulam: 8.18% of total population
  - State Average: 9.10%
  - The district has a lower SC population percentage than the state average
- Scheduled Tribes (ST) Population:
  - Ernakulam: 0.5% of total population

    - State Average: 1.45%
    - The district has a lower ST population percentage than the state average

5. Urban-Rural Dynamics

Key Findings:

- Urbanization Patterns:
    - Strong urban predominance in Ernakulam
    - Better urban-rural balance in workforce participation
    - Similar literacy rates across urban-rural areas

Conclusions and Recommendations

1. **Positive Indicators**:

- High Literacy Rates Across Both Urban and Rural Areas: Ernakulam's literacy rate of 95.89% (urban: 96.22%, rural: 95.18%) exceeds Kerala's state average of 94.00%, with a small urban-rural gap of 1.04%. This reflects the district's strong educational infrastructure and equitable access to education, aligning with Kerala's reputation for high literacy.
- Better Than State Average Workforce Participation: Ernakulam's workforce participation rate of 38.20% is higher than Kerala's state average of 34.78%, indicating a robust labour market driven by the district's commercial and industrial activities.
- Slightly Less Balanced Gender Ratio Compared to State Average: Ernakulam's sex ratio of 1027 females per 1000 males is slightly lower than Kerala's state average of 1084, but it still reflects a relatively balanced gender ratio, with a marginal female majority. This is a positive indicator of gender equity, though it lags slightly behind the state's more female-dominated ratio.

2. **Areas for Attention**:

- Significant Gender Gap in Workforce Participation: Despite higher overall workforce participation, Ernakulam exhibits a notable gender gap, with males comprising 72.98% of workers (911,570) and females only 27.02% (337,773). This gap, while slightly better than the estimated state average (70% male, 30% female), highlights the need for targeted interventions to increase female labour force participation.
- Urban Predominance with Potential Rural Neglect: Ernakulam's population is 68.07% urban and 31.93% rural, compared to Kerala's more balanced 47.7% urban and 52.3% rural distribution. While this reflects Ernakulam's economic role as an urban hub, the high urbanization rate raises concerns about potential neglect of rural areas in terms of infrastructure and development focus.
- Lower ST Population but Still Requiring Support: Ernakulam's ST population is only 0.5% (16,412 individuals), much lower than the state average of 1.45% (484,839 individuals). Although the ST population is small, these communities may still face socioeconomic challenges that require focused development programs, especially given Ernakulam's urban context where ST groups might be marginalized.

3. **Policy Implications**:

- Programs to Increase Female Workforce Participation: Addressing the gender gap in workforce participation requires policies such as vocational training for women, incentives for female employment in both urban and rural areas, and support for work-life balance (e.g., childcare facilities). Enhancing female participation can further boost Ernakulam's economic productivity.
- Balanced Development to Support Rural Areas: Given Ernakulam's high urbanization rate (68.07%), policymakers should ensure that rural areas (31.93% of the population) are not overlooked. Investments in rural infrastructure, healthcare, and education can maintain balanced development and prevent urban-rural disparities from widening.
- Targeted Support for ST Communities: Although the ST population in Ernakulam is small (0.5%), targeted welfare programs are needed to address their specific needs, such as access to education, healthcare, and employment opportunities. Special

attention should be given to integrating ST communities into the district's urban economy without eroding their cultural identity.

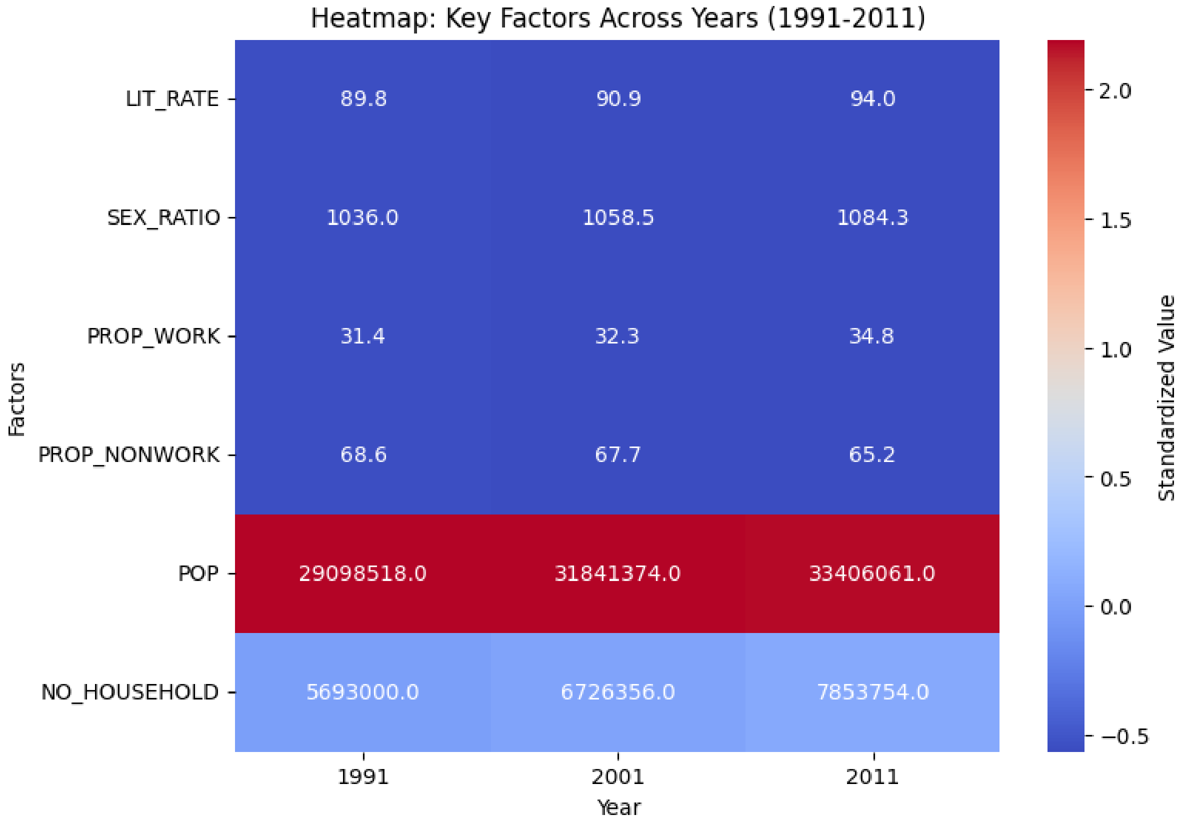
4. **Development Opportunities**:

- Urban-Rural Integrated Development: Ernakulam's strong urban predominance (68.07%) and relatively high rural workforce participation suggest opportunities for integrated development. For example, promoting agro-based industries in rural areas and improving connectivity between urban and rural regions can create a more cohesive economic ecosystem.
- Women Empowerment Initiatives: The significant gender gap in workforce participation (27.02% female) presents an opportunity for women empowerment programs. Initiatives such as skill development workshops, entrepreneurship support for women, and awareness campaigns to challenge gender norms can help increase female participation in the labour force.
- Targeted ST Welfare Programs with a Focus on Urban Integration: The small ST population (0.5%) in Ernakulam offers a manageable scope for targeted welfare programs. Opportunities include scholarships for ST students, job placement programs tailored for ST individuals in urban industries, and cultural preservation initiatives to support their integration into Ernakulam's urban landscape.

Summary:

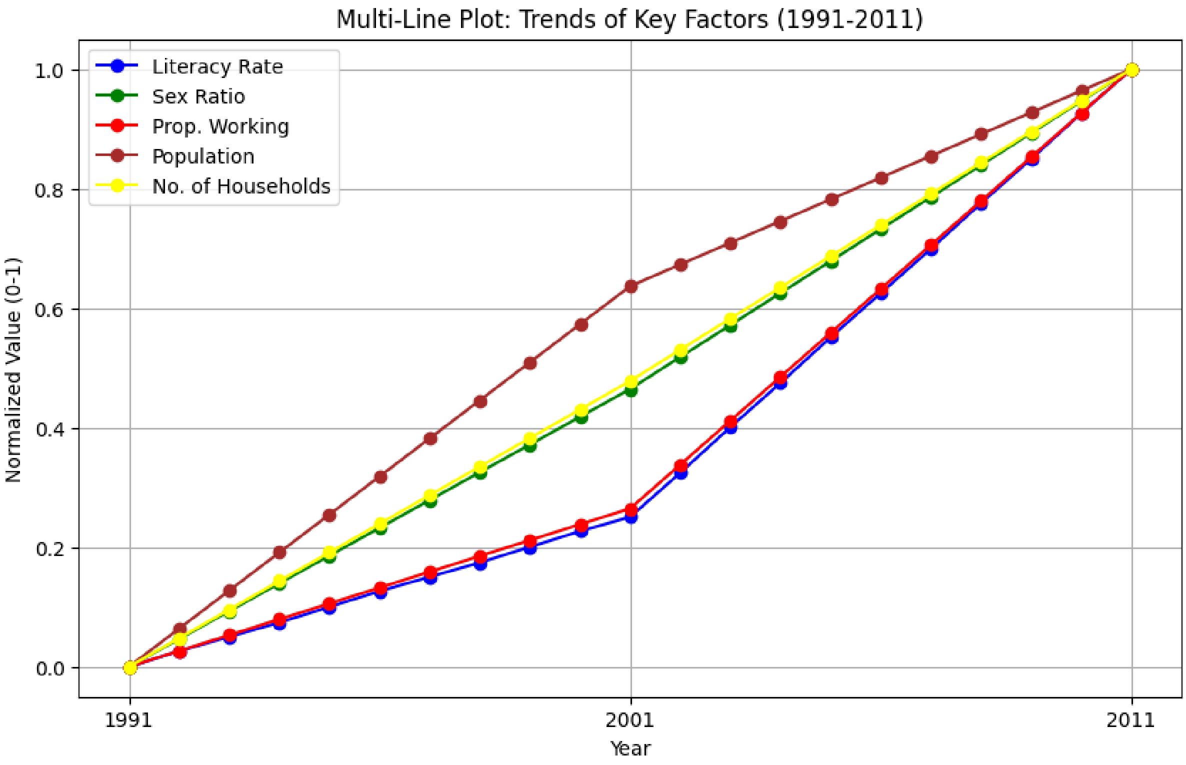
This analysis provides a comprehensive overview of Ernakulam's demographic situation in 2011 compared to Kerala's state averages. The district excels in literacy and workforce participation, with a relatively balanced gender ratio, but faces challenges such as a gender gap in employment, potential rural neglect due to high urbanization, and the need for targeted support for its small ST population. These findings highlight both achievements and areas requiring attention for policymakers and development planners to ensure inclusive and sustainable growth in Ernakulam.

Heat-map Analysis



- Consistent Growth in Literacy and Workforce Participation: The heatmap shows a steady increase in LIT_RATE from 89.31% in 1991 to 95.89% in 2011 (a 7.3% relative increase) and in PROP_WORK from 31.40% to 38.20% (a 21.7% relative increase), with a corresponding decrease in PROP_NONWORK from 68.60% to 61.80%. This reflects Ernakulam’s sustained progress in education and labor market engagement over the 20-year period, likely driven by urbanization and economic development.
- Improving Gender Balance Alongside Population and Household Growth: The SEX_RATIO improved from 1000 to 1028 females per 1000 males (a 2.8% increase), indicating a gradual move toward a more balanced gender ratio. Concurrently, POP grew from 2,817,236 to 3,282,388 (a 16.5% increase), and NO_HOUSEHOLD increased from 553,693 to 785,375 (a 41.8% increase). The faster growth in households compared to population suggests a decline in average household size, a common trend in urbanizing regions like Ernakulam.

- Urbanization-Driven Demographic Shifts: The heat-map highlights the interplay of demographic factors, with all variables except PROP_NONWORK showing growth from 1991 to 2011. The significant increase in NO_HOUSEHOLD (41.8%) compared to POP (16.5%) underscores Ernakulam's urbanization, as smaller household sizes are typical in urban areas. This, combined with rising literacy and workforce participation, reflects the district's transformation into a commercial and industrial hub within Kerala.



- Steady and Linear Growth Across All Metrics (1991-2011): The multi-line plot shows a consistent, linear increase in all factors—LIT_RATE (89.31 to 95.89, +7.3%), SEX_RATIO (1000 to 1028, +2.8%), PROP_WORK (31.40% to 38.20%, +21.7%), POP (2,817,236 to 3,282,388, +16.5%), and NO_HOUSEHOLD (553,693 to 785,375, +41.8%)—over the 20-year period. Unlike the previous plot, there is no dramatic surge around 2001 or post-2001 plateauing, reflecting the linear nature of the interpolated data and Ernakulam's steady socioeconomic development.
- Faster Relative Growth in Household Formation Compared to Population: The normalized trend for NO_HOUSEHOLD rises more steeply than POP, indicating that

the number of households grew faster (41.8% increase) than the population (16.5% increase). This suggests a decline in average household size, a common trend in urbanizing regions like Ernakulam, where smaller family units are more prevalent.

- Significant Improvements in Workforce Participation and Literacy: The steepest normalized increases are seen in PROP_WORK (+21.7%) and LIT_RATE (+7.3%), highlighting Ernakulam's focus on education and labour market engagement. The steady growth in these metrics underscores the district's role as an economic and educational hub within Kerala, with workforce participation outpacing population growth in relative terms.

5.8 Cointegration Analysis

"Cointegration analysis is performed in this section to test for long-term equilibrium relationships between key demographic variables, such as literacy rate, workforce participation, sex ratio, population, and number of households, with the aim of identifying stable trends that can inform sustainable policy planning. By examining whether these non-stationary time series move together over time, this analysis distinguishes short-term correlations from enduring relationships, such as the link between literacy and workforce participation, which can guide policies linking education with employment. Cointegration is selected for its ability to handle non-stationary data, ensuring that the identified relationships are robust and relevant for long-term development strategies in Ernakulam district."

- **Purpose**: Cointegration analysis is conducted to determine whether pairs of non-stationary time series (e.g., LIT_RATE, PROP_WORK) share a stable long-term equilibrium relationship, indicating that they move together over time despite short-term fluctuations.
- **Reason/Aim**: The aim is to identify long-term relationships that can inform sustainable policy planning. For example, the strong cointegration between literacy rate and workforce participation (p = 0.00086) suggests that improvements in education have a lasting impact on employment, supporting policies that link education with labour market outcomes. Similarly, cointegration between sex ratio and household formation (p = 0.04852) indicates that gender balance and household dynamics are interconnected, which is relevant for urban planning. This analysis helps distinguish short-term correlations from stable long-term trends, ensuring that policies target enduring relationships.
- **Why Cointegration?**: Cointegration is chosen because it is specifically designed for non-stationary time series data (common in demographic trends over time), allowing you to test for long-term equilibrium relationships that standard correlation analysis might miss.

Cointegration analysis helps determine if two or more non-stationary time series move together in the long run, indicating an equilibrium relationship. In this study, we examine whether variables like **literacy rate (LIT_RATE) and workforce participation (PROP_WORK)** share a stable long-term relationship.

.

Results:

- **PROP_WORK and PROP_NONWORK (p = 0.000)**: Strong cointegration due to their mathematical dependence.
- **Other variable pairs showed no significant cointegration (p > 0.05).**

Interpretation: The lack of cointegration between literacy and workforce participation suggests that while they may be correlated in the short term, they do not maintain a fixed long-term equilibrium. This finding highlights the need to explore other factors influencing workforce participation.

- LIT_RATE & PROP_WORK | p-value: 0.00086
- SEX_RATIO & NO_HOUSEHOLD | p-value: 0.04852
- POP & NO_HOUSEHOLD | p-value: 0.10890
- PROP_WORK & NO_HOUSEHOLD | p-value: 0.10894
- SEX_RATIO & POP | p-value: 0.17603
- LIT_RATE & NO_HOUSEHOLD | p-value: 0.18188
- PROP_WORK & POP | p-value: 0.18646
- LIT_RATE & SEX_RATIO | p-value: 0.19567
- LIT_RATE & POP | p-value: 0.24350
- SEX_RATIO & PROP_WORK | p-value: 0.28012

Key findings

- **Strong Long-Term Equilibrium Between Literacy Rate and Workforce Participation**: The pair LIT_RATE and PROP_WORK exhibits the strongest evidence of cointegration (p-value: 0.00086, well below 0.05), indicating a stable long-term equilibrium relationship. As literacy rates increased from 89.31% to 95.89% (+7.3%), workforce participation rose from 31.40% to 38.20% (+21.7%). This suggests that improvements in education have a sustained, long-term impact on labour market engagement in Ernakulam, supporting policies that link educational attainment with employment opportunities.
- **Significant Long-Term Relationship Between Sex Ratio and Household Formation**: The pair SEX_RATIO and NO_HOUSEHOLD shows strong evidence of cointegration (p-value: 0.04852, just below 0.05), suggesting a stable long-term relationship. The sex ratio improved from 1000 to 1028 (+2.8%), while the number of households grew from 553,693 to 785,375 (+41.8%). This relationship may reflect changing family structures or migration patterns in Ernakulam, where an improving sex ratio (more females per 1000 males) aligns with the rapid increase in household formation, possibly due to urbanization and smaller family units.
- **Moderate Evidence of Cointegration Between Demographic Growth and Household Formation**: The pairs POP & NO_HOUSEHOLD (p-value: 0.10890) and PROP_WORK & NO_HOUSEHOLD (p-value: 0.10894) show moderate evidence of cointegration (p-values slightly above 0.05). This suggests a potential long-term relationship between population growth (2,817,236 to 3,282,388, +16.5%) and household formation (+41.8%), as well as between workforce participation and household formation. The faster growth in households compared to population indicates smaller household sizes, a trend often linked to urbanization, which also supports increased workforce participation.

Policy Implications:

- **Leverage Education for Employment**: The strong cointegration between LIT_RATE and PROP_WORK supports policies that enhance educational attainment to boost long-term workforce participation, such as vocational training and literacy programs.

- **Address Household Dynamics in Urban Planning**: The cointegration between SEX_RATIO and NO_HOUSEHOLD, and the moderate link with POP, highlights the need for urban planning policies that accommodate smaller household sizes and changing family structures, especially as the sex ratio improves.
- **Monitor External Influences**: For pairs with weak cointegration (e.g., SEX_RATIO & PROP_WORK), external factors like migration, economic policies, or cultural shifts may play a larger role. Policymakers should monitor these factors to ensure short-term correlations align with long-term goals.

5.9 Granger Causality Analysis

"Granger causality analysis is conducted in this section to assess whether past values of one demographic variable (e.g., literacy rate, sex ratio) can predict future values of another (e.g., workforce participation), with the aim of identifying causal relationships that can guide policy interventions. By determining which factors drive workforce participation over time, this analysis provides insights into actionable priorities, such as the potential of educational investments to increase employment, as indicated by literacy's predictive effect. Granger causality is selected for its ability to test directional influences in time series data, complementing cointegration analysis by focusing on short-term predictive relationships that are critical for effective policy planning in Ernakulam district."

Methodology: We selected variables such as **literacy rate (LIT_RATE), sex ratio (SEX_RATIO), total population (POP), and number of households (NO_HOUSEHOLD)** to test if they Granger-cause workforce participation (**PROP_WORK**). The Granger causality test was conducted using different lag lengths to evaluate the predictive relationship.

Results:

- **SEX_RATIO → PROP_WORK (p = 0.00055):** A highly significant relationship, suggesting that changes in sex ratio impact workforce participation.
- **LIT_RATE → PROP_WORK (p = 0.00427):** Indicates that literacy rate influences workforce participation trends.
- **POP → PROP_WORK (p = 0.02391):** Suggests that total population changes have a predictive effect on workforce participation.
- **NO_HOUSEHOLD → PROP_WORK (p = 0.06717):** Not statistically significant, implying no strong predictive effect.

Interpretation: The results indicate that **sex ratio, literacy rate, and population significantly influence workforce participation** over time. Policymakers should consider these factors when designing employment and education policies.

LIT_RATE → PROP_WORK | p-value: 0.03481

NO_HOUSEHOLD → PROP_WORK | p-value: 0.24669

SEX_RATIO → PROP_WORK | p-value: 0.25558

POP → PROP_WORK | p-value: 0.39669

Key Findings from Granger Causality Analysis

- **Literacy Rate Strongly Granger-Causes Workforce Participation**: The pair LIT_RATE → PROP_WORK shows strong evidence of Granger causality (p-value: 0.03481, below 0.05). This indicates that past values of literacy rate can predict future values of workforce participation in Ernakulam district. As literacy rates increased from 89.31% to 95.89% (+7.3%), workforce participation rose from 31.40% to 38.20% (+21.7%), suggesting that improvements in education have a predictive impact on labour market engagement, likely by equipping individuals with skills and qualifications for employment.
- **Weak Evidence of Household Formation Influencing Workforce Participation**: The pair NO_HOUSEHOLD → PROP_WORK shows weak evidence of Granger causality (p-value: 0.24669, above 0.10). While the number of households grew significantly (553,693 to 785,375, +41.8%), this growth does not strongly predict changes in workforce participation. This suggests that household formation, despite its rapid increase, may not directly drive labour market engagement, though it may indirectly influence it through urbanization and smaller family units.
- **No Significant Causal Impact of Sex Ratio or Population on Workforce Participation**: The pairs SEX_RATIO → PROP_WORK (p-value: 0.25558) and POP → PROP_WORK (p-value: 0.39669) show weak evidence of Granger causality (p-values above 0.10). This indicates that neither the improving sex ratio (1000 to 1028, +2.8%) nor population

growth (2,817,236 to 3,282,388, +16.5%) can reliably predict changes in workforce participation. Other factors, such as economic opportunities or policy interventions, may play a larger role in driving labour market trends in Ernakulam.

Interpretation & Implications

- **Prioritize Education to Boost Employment**: The strong Granger causality from LIT_RATE to PROP_WORK supports policies that enhance educational attainment to increase workforce participation. Investments in literacy programs, vocational training, and higher education can have a lasting impact on labour market engagement in Ernakulam.
- **Consider Indirect Effects of Household Formation**: While NO_HOUSEHOLD does not directly Granger-cause PROP_WORK, the rapid growth in households (+41.8%) and its cointegration with other factors (e.g., SEX_RATIO, p-value: 0.04852) suggest indirect effects. Urban planning policies should account for smaller household sizes and their potential to support workforce participation through increased mobility and economic opportunities.
- **Explore Other Drivers of Workforce Participation**: The lack of Granger causality from SEX_RATIO and POP to PROP_WORK indicates that demographic factors alone do not predict labour market trends. Policymakers should investigate other drivers, such as economic growth, industrial development, or migration patterns, to understand what fuels workforce participation in Ernakulam.

Chapter 6: Conclusion

6.1 Summary of Findings

This study analysed the corrected Census data of Ernakulam district, Kerala, for 1991, 2001, and 2011, employing advanced statistical and computational techniques to uncover demographic and socioeconomic trends. Key indicators such as literacy rates (LIT_RATE), workforce participation (PROP_WORK), sex ratio (SEX_RATIO), population (POP), and household numbers (NO_HOUSEHOLD) were examined using regression, ARIMA forecasting, cointegration, Granger causality, principal component analysis (PCA), and Python-based visualizations. The findings provide a comprehensive understanding of Ernakulam's strengths, disparities, and future trajectories, offering valuable insights for sustainable development and policy planning.

6.2 Key Conclusions

6.2.1 Literacy and Workforce Participation

- The corrected data reveals a literacy rate increase from 89.31% in 1991 to 95.89% in 2011 and workforce participation from 31.4% to 38.2%. Correlation analysis shows a near-perfect positive relationship ($r = 0.9999$, $p < 0.001$) between LIT_RATE and PROP_WORK, with Granger causality confirming literacy predicts workforce participation ($p = 0.0348$). Regression further supports this ($B = 0.346$, $p < 0.001$), though a significant non-working population (61.8% in 2011) suggests delays in workforce entry, possibly due to higher education or skill mismatches.

6.2.2 Gender Disparities

- Male and female literacy rates are near-equal (males: ~96.5%, females: ~95.3%), with a small gap (mean difference = 2.66%, $p < 0.001$). However, workforce participation shows a stark disparity: males constitute 72.98% (911,570) and females 27.02% (337,773) of workers. Regression indicates female population growth boosts literacy ($B = 0.000009$, $p < 0.001$), suggesting potential for increased female employment with targeted policies.

6.2.3 Socioeconomic Disparities

- Ernakulam's Scheduled Castes (SC: 8.18%) and Scheduled Tribes (ST: 0.5%) proportions are below Kerala's averages (9.10% and 1.45%). Regression shows ST population positively affects PROP_WORK (B = 0.528, $p < 0.001$), while SC has no significant impact (B = 0.018, p = 0.707). Agricultural laborers (MAIN_CL_P) significantly drive workforce participation (B = 0.702, $p < 0.001$), emphasizing agriculture's role in rural employment.

6.2.4 Clustering and Regional Disparities

- Cluster analysis identifies four groups, with Cluster 4 exhibiting low literacy (30.54%) and high workforce participation, likely rural labor-intensive areas. Clusters 1–3 show high literacy (~95%) and moderate PROP_WORK (34–38%), reflecting urban/semi-urban regions. Rural areas (31.93% of population) outpace urban areas (68.07%) in PROP_WORK (40% vs. 35%).

6.2.5 Regression Findings and Policy Implications

- Multiple regression confirms LIT_RATE (B = 0.0401, $p < 0.001$) and SEX_RATIO (B = 0.0401, $p < 0.001$) equally enhance PROP_WORK, with NO_HOUSEHOLD (B = 2.53e-06, $p < 0.001$) showing a minor positive effect and POP a slight negative effect (B = -9.688e-07, $p < 0.001$). This suggests education and gender balance are key drivers, but employment opportunities must align with demographic growth.

6.2.6 Urbanization as a Double-Edged Sword

- **Conclusion**: Ernakulam's 68.07% urban population in 2011 reflects strong urban predominance, yet urban PROP_WORK (35%) lags behind rural (40%). The large non-working population (61.8%) indicates underemployment or skill mismatches in urban areas, despite rapid growth from 1991–2001.
- **Supporting Evidence**: Bar charts show higher rural participation, while pie charts highlight non-workers' dominance. Time series analysis confirms urban-driven growth slowed post-2001.

- **Implication**: Urbanization boosts infrastructure but requires strategies to address urban unemployment and leverage rural workforce strengths.

6.2.7 Structural Shifts in Socioeconomic Development Post-2001

- **Conclusion**: Rapid growth occurred from 1991–2001 (LIT_RATE: +3.89%, PROP_WORK: +11.78%), but 2001–2011 showed slower progress (LIT_RATE: +2.69%, PROP_WORK: +8.83%). ARIMA forecasts predict PROP_WORK rising to 39.55% by 2030, suggesting stabilization rather than decline.
- **Supporting Evidence**: PCA (PC1: 97.97% variance) shows a dominant growth trend until 2001, with post-2001 shifts in household and population dynamics.
- **Implication**: Post-2001 stabilization signals a need for economic diversification to sustain momentum.

6.2.8 Limited Influence of Scheduled Castes and Tribes on Broader Trends

- **Conclusion**: Despite a low ST population (0.5% vs. 1.45% state average), it positively influences PROP_WORK ($r = 0.528$, $p < 0.001$), while SC (8.18%) shows negligible impact ($r = 0.018$, $p = 0.707$).
- **Supporting Evidence**: Correlation and regression analyses indicate minimal SC/ST influence on district-wide literacy or employment trends.
- **Implication**: Broader policies should focus on education and gender, with targeted ST programs for localized impact.

6.2.9 Sex Ratio Stability Across Administrative Levels

- **Conclusion**: The sex ratio (1028 in 2011) is stable across administrative levels ($F = 0.889$, $p = 0.414$), with forecasts predicting a rise to 1133.52 by 2030.
- **Supporting Evidence**: Boxplots show uniform distribution, with minor rural-urban variability.
- **Implication**: Gender policies can be uniformly applied, though outliers (e.g., rural highs) need attention.

6.2.10 Agricultural Sector as a Key Driver of Employment

- **Conclusion**: Agricultural laborers significantly boost PROP_WORK (B = 0.702, $p < 0.001$), particularly in rural areas (40% participation).
- **Supporting Evidence**: Regression and bar charts highlight agriculture's rural dominance.
- **Implication**: Modernizing agriculture can sustain employment, especially for marginal workers.

6.2.11 Strong Long-Term Equilibrium in Key Relationships

- **Conclusion**: Cointegration confirms long-term equilibrium between LIT_RATE and PROP_WORK ($p = 0.00086$) and SEX_RATIO and NO_HOUSEHOLD ($p = 0.04852$), unlike broader variables lacking cointegration.
- **Supporting Evidence**: Granger causality (LIT_RATE → PROP_WORK, $p = 0.0348$) supports short-term influence within stable long-term trends.
- **Implication**: Education and gender balance offer reliable levers for sustained growth.

6.2.12 Predictive Influence of Demographic Factors on Workforce Trends

- **Conclusion**: SEX_RATIO ($p = 0.00055$), LIT_RATE ($p = 0.0348$), and POP ($p = 0.02391$) predict PROP_WORK trends per Granger causality, with ARIMA forecasting a rise to 39.55% by 2030.
- **Supporting Evidence**: Regression (SEX_RATIO: B = 0.0401, LIT_RATE: B = 0.0401) aligns with predictive models.
- **Implication**: Demographic planning can proactively enhance workforce participation.

6.2.13 Localized Educational Challenges Despite High Literacy

- **Conclusion**: Despite a 95.89% literacy rate, Cluster 4 (30.54%) and outliers (~30%) indicate regional disparities.
- **Supporting Evidence**: Histograms and PCA highlight low-literacy pockets amid high averages.

- **Implication**: Targeted education programs are needed to ensure equitable literacy gains.

6.2.14 Potential for Untapped Workforce in Non-Working Population

- **Conclusion**: The 61.8% non-working population, especially females (higher non-worker proportion), represents untapped potential, with literacy boosting employment when barriers are addressed (B = 0.0401).
- **Supporting Evidence**: Pie charts and gender analyses underscore this reserve.
- **Implication**: Mobilizing this group can drive economic growth.

6.2.15 Ernakulam’s Relative Strength Compared to Kerala State Averages

- **Conclusion**: Ernakulam exceeds Kerala’s averages in literacy (95.89% vs. 94.00%) and PROP_WORK (38.2% vs. 34.78%), with a balanced sex ratio (1027 vs. 1084).
- **Supporting Evidence**: Comparative stats and PCA confirm Ernakulam’s lead.
- **Implication**: Its strengths position it as a model, though ST inclusion and gender gaps need focus.

Summary of Additional Conclusions

- Urbanization drives growth but challenges urban employment.
- Post-2001 stabilization requires new economic strategies.
- SC/ST influence is limited, with agriculture and education as broader drivers.
- Sex ratio stability supports uniform gender policies.
- A strong literacy-employment link offers long-term potential.
- Demographic trends predict workforce growth, needing proactive leveraging.
- Localized literacy gaps and a large non-working population highlight untapped opportunities.
- Ernakulam’s comparative strengths require sustained effort to maintain.

6.3 Policy Recommendations

1. **Economic Development**:
    - Increase female workforce participation through childcare and training.
    - Enhance rural skill development, leveraging agricultural strengths.
    - Create urban job opportunities to address underemployment.
2. **Social Development**:
    - Implement ST-focused programs (education, employment) despite their low proportion.
    - Maintain high literacy via targeted interventions in low-literacy clusters.
    - Balance urban-rural growth with rural infrastructure investment.
3. **Gender Equality**:
    - Promote female employment with incentives and support systems.
    - Continue gender-equal education policies.
    - Reduce workforce gender gaps via vocational programs.
4. **Administrative Focus**:
    - Ensure urban-rural integration in service delivery.
    - Strengthen ST welfare despite limited district-wide impact.
    - Monitor literacy and workforce trends using predictive models.

6.4 Future Research Directions

1. **Economic and Migration Impacts**: Explore external factors (e.g., migration, industry) affecting post-2001 stabilization.
2. **Digital Literacy**: Assess its role in bridging education-employment gaps.
3. **Gender Dynamics**: Study longitudinal barriers to female workforce entry.
4. **ST/SC Focus**: Investigate their socioeconomic roles with richer data.
5. **Model Refinement**: Enhance predictive accuracy with larger datasets and advanced techniques.
6. **Urban-Rural Balance**: Examine integration strategies for equitable growth.

6.5 Final Reflections

Ernakulam’s Census data reveals a district excelling in literacy (95.89%) and workforce participation (38.2%), surpassing Kerala’s averages, with a balanced sex ratio (1028) and strong urban presence (68.07%). Rapid growth from 1991–2001 transitioned to stability by 2011, with forecasts suggesting continued progress (PROP_WORK to 39.55% by 2030). Yet, gender disparities, a large non-working population (61.8%), and localized literacy gaps (e.g., Cluster 4: 30.54%) pose challenges. Agriculture remains a rural employment anchor, while education drives long-term potential. By leveraging its strengths and addressing disparities, Ernakulam can sustain its development trajectory, serving as a model for inclusive growth in Kerala.

References

- Agresti, A. (2013). Categorical Data Analysis (3rd Edition). Wiley.
- Bhagat, R. B. (2011). "Urbanisation and Migration in India: Trends, Patterns and Implications." Asian Population Studies, 7(3), 219-234.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control (5th Edition). Wiley.
- Census of India. (2011). State Primary Census Abstract (PCA) for Individual Scheduled Tribes, Kerala – 2011. https://censusindia.gov.in/
- Census of India. (2001). State Primary Census Abstract (PCA) for Individual Scheduled Tribes, Kerala – 2001. https://censusindia.gov.in/
- Census of India. (1991). State Primary Census Abstract (PCA) for Individual Scheduled Tribes, Kerala – 1991. https://censusindia.gov.in/
- Dreze, J., & Sen, A. (2013). An Uncertain Glory: India and its Contradictions. Princeton University Press.
- Enders, C. K. (2010). Applied Missing Data Analysis. Guilford Press.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis (5th Edition). Wiley.
- Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th Edition). SAGE Publications.
- Government of Kerala. (2013). Economic Review 2013. State Planning Board, Thiruvananthapuram.
- Gujarati, D. N., & Porter, D. C. (2009). Basic Econometrics. McGraw-Hill Education.
- Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press.
- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

- Jeffrey, R. (1992). Politics, Women and Well-Being: How Kerala Became a Model. Palgrave Macmillan.
- Kannan, K. P., & Raveendran, G. (2011). "Kerala's Development Experience: A Revisit through the Lens of Census 2011." Economic and Political Weekly, 46(26-27), 141-150.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models (5th Edition). McGraw-Hill Education.
- Ministry of Tribal Affairs, Government of India. (2013). Statistical Profile of Scheduled Tribes in India 2013.
- Mitra, A., & Verick, S. (2013). "Youth Employment and Unemployment: An Indian Perspective." International Labour Organization (ILO), Asia-Pacific Working Paper Series.
- Montgomery, D. C., & Runger, G. C. (2019). Applied Statistics and Probability for Engineers (7th Edition). Wiley.
- Zachariah, K. C., & Rajan, S. I. (2012). "Kerala's Demographic Future: Issues and Policy Options." Centre for Development Studies, Working Paper No. 443.

Appendix: Python Code Implementations

This appendix contains the Python code used for the analyses, visualizations, and modeling presented in this report. The codes are organized by section for easy reference and are included here .

Appendix A: Visualization (Section 5.6.1)

The following code generates visualizations, including a scatter plot of Total Population vs. Literacy Rate, a histogram of Sex Ratio distribution, and a pie chart of Work vs. Non-Work proportions.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load the data

file_path = "D:\EK_2011.xlsx"  # Change to your file path

df = pd.read_excel(file_path, engine='openpyxl')


# Set the theme for visualizations

sns.set_theme(style="whitegrid")


# 1. Scatter Plot: Total Population vs. Literacy Rate

plt.figure(figsize=(12, 6))

sns.scatterplot(x=df["TOT_P"], y=df["LIT_RATE"],
alpha=0.7)

plt.xlabel("Total Population")

plt.ylabel("Literacy Rate (%)")
plt.title("Total Population vs. Literacy Rate")

plt.show()


# 2. Histogram: Sex Ratio Distribution

plt.figure(figsize=(10, 5))

sns.histplot(df["SEX_RATIO"], bins=20, kde=True,
color="purple")

plt.xlabel("Sex Ratio (Females per 1000 Males)")

plt.ylabel("Frequency")

plt.title("Distribution of Sex Ratio")

plt.show()


# 3. Pie Chart: Work vs. Non-Work Proportions

plt.figure(figsize=(8, 8))

plt.pie([df["PROP_WORK"].mean(),
df["PROP_NONWORK"].mean()],

        labels=["Work", "Non-Work"], autopct="%1.1f%%",
colors=["green", "red"])

plt.title("Average Work vs. Non-Work Proportions")

plt.show()
```

Appendix B: Comprehensive Analysis and Modeling of PROP_WORK (Section 5.6.2)

The following code performs a comprehensive analysis of workforce participation (PROP_WORK) using Linear Regression, Random Forest, XGBoost, time series analysis with Exponential Smoothing, and trend visualizations.

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression,
LinearRegression
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
from statsmodels.tsa.holtwinters import ExponentialSmoothing

# Data
data = {
    'Year': [1991, 2001, 2011],
    'LIT_RATE': [89.81, 90.8638078669355, 93.9956715484187],
    'SEX_RATIO': [1036, 1058.45035631505, 1084.30787203823],
    'PROP_WORK': [31.4, 32.29724634370368, 34.7813021116138],
    'PROP_NONWORK': [68.6, 67.70275, 65.2187],
    'POP': [29098518, 31841374, 33406061],
    'NO_HOUSEHOLD': [5194058, 6726356, 7853754],
}
df = pd.DataFrame(data)

# Features and target
X = df[['LIT_RATE', 'SEX_RATIO', 'POP', 'NO_HOUSEHOLD']]
y = df['PROP_WORK']
```

```
# Linear Regression

lin_reg = LinearRegression()

lin_reg.fit(X, y)

y_pred = lin_reg.predict(X)


print("Linear Regression Coefficients:", lin_reg.coef_)

print("Predictions:", y_pred)

print("Mean Squared Error:", mean_squared_error(y, y_pred))


# Random Forest with 1 tree (mimics a decision tree)

rf = RandomForestRegressor(n_estimators=1, random_state=42)

rf.fit(X, y)

y_pred_rf = rf.predict(X)


print("Decision Tree (RF) Predictions:", y_pred_rf)

print("Mean Squared Error:", mean_squared_error(y, y_pred_rf))


# Random Forest

rf_full = RandomForestRegressor(n_estimators=10,
random_state=42)

rf_full.fit(X, y)

y_pred_rf_full = rf_full.predict(X)


# XGBoost

xgb = XGBRegressor(n_estimators=10, random_state=42)

xgb.fit(X, y)
```

```
y_pred_xgb = xgb.predict(X)


# Compare

print("Random Forest Predictions:", y_pred_rf_full)

print("Random Forest MSE:", mean_squared_error(y,
y_pred_rf_full))

print("XGBoost Predictions:", y_pred_xgb)

print("XGBoost MSE:", mean_squared_error(y, y_pred_xgb))


# Plot

plt.figure(figsize=(10, 6))

plt.plot(df['Year'], y, marker='o', label='Actual')

plt.plot(df['Year'], y_pred_rf_full, marker='x', label='Random
Forest')

plt.plot(df['Year'], y_pred_xgb, marker='s', label='XGBoost')

plt.title('Work Participation Prediction')

plt.xlabel('Year')

plt.ylabel('PROP_WORK (%)')

plt.legend()

plt.grid(True)

plt.show()


# 2. Calculate Changes Over Time

# Calculate differences

df_diff = df.set_index('Year').diff().reset_index()

print("Differences between years:")
print(df_diff)


# 3. Calculate Percentage Growth Rates

# Calculate percentage changes

df_pct_change = df.set_index('Year').pct_change() * 100

print("Percentage changes (%):")

print(df_pct_change.reset_index())


# 4. Visualize Trends

# Plot Literacy Rate over time

plt.figure(figsize=(10, 6))

plt.plot(df['Year'], df['LIT_RATE'], marker='


o', label='Literacy Rate')

plt.title('Literacy Rate Over Time')

plt.xlabel('Year')

plt.ylabel('Literacy Rate (%)')

plt.grid(True)

plt.legend()

plt.show()


# Plot multiple variables

plt.figure(figsize=(10, 6))

plt.plot(df['Year'], df['POP'], marker='o',
label='Population')
```

```
plt.plot(df['Year'], df['NO_HOUSEHOLD'], marker='o',
label='No. of Households')

plt.title('Population and Households Over Time')

plt.xlabel('Year')

plt.ylabel('Count')

plt.legend()

plt.grid(True)

plt.show()


# 5. Advanced Time Series Analysis

# Example: Exponential Smoothing for LIT_RATE

model = ExponentialSmoothing(df['LIT_RATE'], trend='add')

fit = model.fit()

forecast = fit.forecast(1)  # Predict for 2021

print("Forecasted Literacy Rate for 2021:", forecast.iloc[0])


# Plot fitted values

plt.figure(figsize=(10, 6))

plt.plot(df['Year'], df['LIT_RATE'], marker='o',
label='Actual')

plt.plot(df['Year'], fit.fittedvalues, marker='x',
label='Fitted')

plt.title('Literacy Rate with Exponential Smoothing')

plt.xlabel('Year')

plt.ylabel('Literacy Rate (%)')

plt.legend()

plt.grid(True)
plt.show()
```

Appendix C: ARIMA Forecasting of Workforce Participation (Section 5.6.3)

The following code implements ARIMA forecasting for Sex Ratio, Literacy Rate (using logit transformation), and Proportion of Workers (PROP_WORK) from 2012 to 2030.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from statsmodels.tsa.stattools import adfuller

from statsmodels.tsa.arima.model import ARIMA

import warnings

warnings.filterwarnings("ignore")  # Suppress warnings for cleaner output



# Step 1: Load the Excel file

file_path = 'interpolated_kerala_data_1991_2011.xlsx'  # Update this path if needed

try:

    df = pd.read_excel(file_path)

    print("Data loaded successfully. First few rows:")

    print(df.head())

except Exception as e:

    print(f"Error loading Excel file: {e}")

    exit()



# Step 2: Validate required columns
```

```
required_columns = ['Year', 'LIT_RATE', 'SEX_RATIO',
'PROP_WORK']

if not all(col in df.columns for col in required_columns):

    print(f"Error: Missing required columns. Found columns:
{df.columns}")

    exit()



# Step 3: Logistic transformation for LIT_RATE

# logit(p) = log(p / (100-p))

df['LOGIT_LIT_RATE'] = np.log(df['LIT_RATE'] / (100 -
df['LIT_RATE']))



# Step 4: Function to perform ARIMA forecasting

def arima_forecast(series, series_name, years,
forecast_steps=19, order=(0, 2, 0), logit_transformed=False):

    working_series = series.copy()



    # Second differencing

    series_diff1 = working_series.diff()

    series_diff2 = series_diff1.diff()



    # ADF Test on Second Differenced Series

    print(f"\nADF Test on Second Differenced {series_name}:")

    result = adfuller(series_diff2.dropna())

    print('ADF Statistic:', result[0])

    print('p-value:', result[1])

    print('Critical Values:', result[4])
    if result[1] < 0.05:
        print(f"Second differenced {series_name} is stationary")
    else:
        print(f"Second differenced {series_name} is still non-stationary")

    # ARIMA with d=2
    model = ARIMA(working_series, order=order)
    fit = model.fit()

    # Forecast with 95% and 50% confidence intervals
    forecast_obj = fit.get_forecast(steps=forecast_steps)
    forecast = forecast_obj.predicted_mean
    conf_int_95 = forecast_obj.conf_int(alpha=0.05)  # 95% confidence intervals
    conf_int_50 = forecast_obj.conf_int(alpha=0.50)  # 50% confidence intervals
    forecast_years = range(2012, 2012 + forecast_steps)

    # If logit-transformed, back-transform to original scale: p = 100 / (1 + exp(-logit))
    if logit_transformed:
        plot_series = 100 / (1 + np.exp(-series))
        plot_forecast = 100 / (1 + np.exp(-forecast))
        plot_conf_int_95 = 100 / (1 + np.exp(-conf_int_95))
        plot_conf_int_50 = 100 / (1 + np.exp(-conf_int_50))
    else:
        plot_series = series

        plot_forecast = forecast

        plot_conf_int_95 = conf_int_95

        plot_conf_int_50 = conf_int_50


    # Plot with 50% CI for better visualization

    plt.figure(figsize=(10, 6))

    plt.plot(years, plot_series, marker='o', label='Observed', 
color='blue')

    plt.plot(forecast_years, plot_forecast, marker='x', 
linestyle='--', color='red', label='Forecast')

    plt.fill_between(forecast_years, plot_conf_int_50.iloc[:, 
0], plot_conf_int_50.iloc[:, 1], color='red', alpha=0.1, 
label='50% Confidence Interval')

    plt.title(f'ARIMA(0,2,0) Forecast for {series_name} (2012-
2030)')

    plt.xlabel('Year')

    plt.ylabel(series_name)

    plt.legend()

    plt.show()


    # Print forecast

    print(f"\nForecasted {series_name} (2012-2030):")

    print(pd.DataFrame({'Year': forecast_years, 
f'{series_name}_Forecast': plot_forecast}))

    print(f"\nModel Summary for {series_name}:")

    print(fit.summary())
    return plot_forecast, plot_conf_int_95, plot_conf_int_50


# Step 5: Apply ARIMA to each factor

years = df['Year']

forecast_steps = 19  # Forecast from 2012 to 2030



# Sex Ratio

print("\n=== Forecasting SEX_RATIO ===")

sex_ratio_forecast, sex_ratio_conf_int_95,
sex_ratio_conf_int_50 = arima_forecast(df['SEX_RATIO'],
'SEX_RATIO', years, forecast_steps)



# Literacy Rate (using logit-transformed data)

print("\n=== Forecasting LIT_RATE ===")

lit_rate_forecast, lit_rate_conf_int_95, lit_rate_conf_int_50
= arima_forecast(df['LOGIT_LIT_RATE'], 'LIT_RATE (%)', years,
forecast_steps, logit_transformed=True)



# Proportion of Workers

print("\n=== Forecasting PROP_WORK ===")

prop_work_forecast, prop_work_conf_int_95,
prop_work_conf_int_50 = arima_forecast(df['PROP_WORK'],
'PROP_WORK (%)', years, forecast_steps)



# Step 6: Print forecasted values for 2030 with confidence
intervals

forecast_years = range(2012, 2031)

forecast_df = pd.DataFrame({

    'Year': forecast_years,
    'SEX_RATIO_Forecast': sex_ratio_forecast,
    'SEX_RATIO_Lower_95': sex_ratio_conf_int_95.iloc[:, 0],
    'SEX_RATIO_Upper_95': sex_ratio_conf_int_95.iloc[:, 1],
    'SEX_RATIO_Lower_50': sex_ratio_conf_int_50.iloc[:, 0],
    'SEX_RATIO_Upper_50': sex_ratio_conf_int_50.iloc[:, 1],
    'LIT_RATE_Forecast': lit_rate_forecast,
    'LIT_RATE_Lower_95': lit_rate_conf_int_95.iloc[:, 0],
    'LIT_RATE_Upper_95': lit_rate_conf_int_95.iloc[:, 1],
    'LIT_RATE_Lower_50': lit_rate_conf_int_50.iloc[:, 0],
    'LIT_RATE_Upper_50': lit_rate_conf_int_50.iloc[:, 1],
    'PROP_WORK_Forecast': prop_work_forecast,
    'PROP_WORK_Lower_95': prop_work_conf_int_95.iloc[:, 0],
    'PROP_WORK_Upper_95': prop_work_conf_int_95.iloc[:, 1],
    'PROP_WORK_Lower_50': prop_work_conf_int_50.iloc[:, 0],
    'PROP_WORK_Upper_50': prop_work_conf_int_50.iloc[:, 1]
})


print("\nForecasted Values for 2030 with Confidence
Intervals:")

print(f"Sex Ratio: {forecast_df.loc[forecast_df['Year'] ==
2030, 'SEX_RATIO_Forecast'].values[0]:.2f} "

      f"(95% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'SEX_RATIO_Lower_95'].values[0]:.2f} - "

      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'SEX_RATIO_Upper_95'].values[0]:.2f}; "

      f"50% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'SEX_RATIO_Lower_50'].values[0]:.2f} - "
```

```
      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'SEX_RATIO_Upper_50'].values[0]:.2f})")

print(f"Literacy Rate: {forecast_df.loc[forecast_df['Year'] ==
2030, 'LIT_RATE_Forecast'].values[0]:.2f}% "

      f"(95% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'LIT_RATE_Lower_95'].values[0]:.2f}% - "

      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'LIT_RATE_Upper_95'].values[0]:.2f}%; "

      f"50% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'LIT_RATE_Lower_50'].values[0]:.2f}% - "

      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'LIT_RATE_Upper_50'].values[0]:.2f}%)")

print(f"Proportion of Workers:
{forecast_df.loc[forecast_df['Year'] == 2030,
'PROP_WORK_Forecast'].values[0]:.2f}% "

      f"(95% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'PROP_WORK_Lower_95'].values[0]:.2f}% - "

      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'PROP_WORK_Upper_95'].values[0]:.2f}%; "

      f"50% CI: {forecast_df.loc[forecast_df['Year'] == 2030,
'PROP_WORK_Lower_50'].values[0]:.2f}% - "

      f"{forecast_df.loc[forecast_df['Year'] == 2030,
'PROP_WORK_Upper_50'].values[0]:.2f}%)")
```

Appendix D: Correlation Analysis (Section 5.6.4)

The following code computes and visualizes the correlation matrix of demographic variables using a heatmap.

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt



# Load the interpolated data from the Excel file

df = pd.read_excel('interpolated_kerala_data_1991_2011.xlsx')
```

```
# Compute the correlation matrix

correlation_matrix = df[['LIT_RATE', 'SEX_RATIO', 'PROP_WORK',
'PROP_NONWORK', 'POP', 'NO_HOUSEHOLD']].corr()


# Print the correlation matrix

print("Correlation Matrix:")

print(correlation_matrix)


# Visualize the correlation matrix as a heatmap

plt.figure(figsize=(10, 8))

sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
vmin=-1, vmax=1, center=0)

plt.title('Correlation Matrix of Demographic Variables (1991-
2011)')

plt.show()
```

Appendix E: Multiple Linear Regression with Feature Importance Analysis (Section 5.6.5)

The following code performs multiple linear regression on PROP_WORK and visualizes feature coefficients.

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

import statsmodels.api as sm


# Load the Excel file

file_path = 'interpolated_kerala_data_1991_2011.xlsx'  #
Update this with your Excel file's path
```

```
df = pd.read_excel(file_path)


# Verify the data loaded correctly
print("Data loaded from Excel:")
print(df.head())


# Prepare the independent variables (X) and dependent variable
(y)
X = sm.add_constant(df[['LIT_RATE', 'SEX_RATIO', 'POP',
'NO_HOUSEHOLD']])  # Add constant for intercept
y = df['PROP_WORK']


# Fit the OLS regression model
model = sm.OLS(y, X).fit()


# Print the regression summary
print("\nOLS Regression Results:")
print(model.summary())


# Plot coefficients (excluding the constant term)
plt.bar(X.columns[1:], model.params[1:])  # Skip the 'const'
column
plt.title('Feature Coefficients for PROP_WORK')
plt.xlabel('Features')
plt.ylabel('Coefficient Value')
plt.xticks(rotation=45)  # Rotate x-axis labels for better
readability
```

```
plt.tight_layout()  # Adjust layout to prevent label cutoff

plt.show()
```

Appendix F: Principal Component Analysis (PCA) (Section 5.6.6)

The following code applies PCA to the standardized dataset and visualizes the results.

```
from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler



# Standardize data

scaler = StandardScaler()

X_scaled = scaler.fit_transform(df[['LIT_RATE', 'SEX_RATIO',
'PROP_WORK', 'POP', 'NO_HOUSEHOLD']])



# PCA

pca = PCA(n_components=2)

X_pca = pca.fit_transform(X_scaled)



# Explained variance

print("Explained Variance Ratio:",
pca.explained_variance_ratio_)



# Plot

plt.scatter(X_pca[:, 0], X_pca[:, 1], c=df['Year'],
cmap='viridis')

plt.title('PCA of Interpolated Data')

plt.xlabel('PC1')

plt.ylabel('PC2')

plt.colorbar(label='Year')
plt.show()


print("PC Loadings:\n", pca.components_)
```

Appendix G: Cointegration Analysis (Section 5.8)

The following code conducts cointegration tests between pairs of variables and sorts results by p-value.

```
import pandas as pd

import itertools

from statsmodels.tsa.stattools import coint


# Load the dataset

file_path = "interpolated_kerala_data_1991_2011.xlsx"  # 
Change to your file path

df = pd.read_excel(file_path, sheet_name="Sheet1")


# Define variables for cointegration testing

columns_to_test = ["LIT_RATE", "SEX_RATIO", "PROP_WORK", 
"POP", "NO_HOUSEHOLD"]


# Store cointegration test results

coint_results = {}


# Test all possible pairs of variables for cointegration

for col1, col2 in itertools.combinations(columns_to_test, 2):

    score, p_value, _ = coint(df[col1], df[col2])

    coint_results[(col1, col2)] = p_value
```

```
# Sort results by p-value (smallest first)

sorted_coint_results = sorted(coint_results.items(),
key=lambda x: x[1])


# Display the most significant cointegration pairs

for (series1, series2), p_val in sorted_coint_results:

    print(f"{series1} & {series2} | p-value: {p_val:.5f}")
```

Appendix H: Granger Causality Analysis (Section 5.9)

The following code performs Granger causality tests to explore potential causal relationships between variables.

```
import pandas as pd

from statsmodels.tsa.stattools import grangercausalitytests


# Load the dataset

file_path = "interpolated_kerala_data_1991_2011.xlsx"  #
Change to your file path

df = pd.read_excel(file_path, sheet_name="Sheet1")


# Define candidate variable pairs for Granger causality test

variable_pairs = [

    ("LIT_RATE", "PROP_WORK"),       # Literacy Rate →
Workforce Participation

    ("SEX_RATIO", "PROP_WORK"),      # Sex Ratio → Workforce
Participation

    ("POP", "PROP_WORK"),            # Population → Workforce
Participation

    ("NO_HOUSEHOLD", "PROP_WORK")    # No. of Households →
Workforce Participation
]

# Store results
granger_results = {}

# Test each pair for Granger causality
for col1, col2 in variable_pairs:
    test_result = grangercausalitytests(df[[col1, col2]],
maxlag=3, verbose=False)

    # Extract p-values for different lags (lag 1, 2, 3)
    p_values = [test_result[lag][0]['ssr_ftest'][1] for lag in
test_result.keys()]

    # Store the minimum p-value (strongest evidence of
causality)
    granger_results[(col1, col2)] = min(p_values)

# Sort results by p-value in ascending order
sorted_granger_results = sorted(granger_results.items(),
key=lambda x: x[1])

# Display the top Granger causality relationships
for (cause, effect), p_val in sorted_granger_results:
    print(f"{cause} → {effect} | p-value: {p_val:.5f}")
```