INTRODUCTION

India, with its vast geographical expanse and socio-cultural diversity, presents a complex landscape of development where progress is often unevenly distributed across regions. Among the various indicators of human development, **housing quality** stands out as a vital measure. It not only provides a sense of security and dignity but also significantly influences key aspects of human well-being such as health outcomes, educational opportunities, and overall economic productivity. In rural and urban contexts alike, the quality and availability of housing infrastructure mirror the level of access to basic services such as water, sanitation, electricity, and structural safety.

Understanding the **spatial and demographic distribution** of housing quality is critical for assessing regional disparities, identifying pockets of underdevelopment, and guiding effective public policy. However, housing conditions alone do not tell the full story. Their interplay with socio-economic indicators such as **literacy rates, female literacy, Scheduled Caste (SC) and Scheduled Tribe (ST) population proportions, population density**, and household structures provides a richer context for understanding the development challenges faced by different districts.

To address this multifaceted issue, this project undertakes a comprehensive data-driven analysis using datasets from the **2011 Indian Census**—one of the most granular and reliable sources of demographic and housing information in the country. The core objective is to explore the relationships between **housing quality** and **socio-demographic characteristics** at the district level, uncovering patterns that are often hidden in aggregated national-level statistics.

A notable innovation introduced in this study is the creation of a **Housing Quality Index (HQI)**. This index is a composite metric, normalized on a 0 to 1 scale, that reflects the condition of housing in each district based on the proportion of houses classified as *good*, *livable*, or *dilapidated*. This allows for consistent comparison across districts and helps in identifying outliers and clusters of poor housing.

The project further incorporates a suite of **modern data science methodologies**:

- **Exploratory Data Analysis (EDA)** to gain descriptive insights into housing and demographic patterns.
- **Feature Engineering** to derive meaningful metrics and indicators from raw data.
- **Correlation Analysis** to quantify the strength of associations between housing quality and other variables.
- **Clustering Algorithms** such as *k-means* to group districts based on similar profiles.
- **Spatial Mapping** using static and interactive maps to visualize regional disparities.
- **Hypothesis Testing and Regression Modeling** to assess statistically significant relationships,

such as the impact of SC/ST population proportions on female literacy rates.

These analytical techniques are implemented using the **R programming language**, an open-source platform well-suited for statistical computing and data visualization. Packages such as tidyverse, ggplot2, sf, leaflet, and cluster are employed to ensure reproducibility, flexibility, and scalability in the analysis.

The findings of this study are not only academic in nature but also **policy-relevant**. They help in identifying structurally disadvantaged regions, interpreting the links between housing and social inequality, and presenting results in a format that is intuitive and actionable for planners, researchers, and government agencies. By highlighting regional imbalances and their potential causes, the project contributes to the broader dialogue on **inclusive development** and **evidence-based policymaking** in India.

Ultimately, this project showcases the power of **data science in the social development domain**, demonstrating how publicly available datasets, when analyzed systematically, can yield valuable insights for improving lives and reducing inequalities at the grassroots level.

OBJECTIVES

The overarching aim of this project is to perform an in-depth analysis of housing quality across Indian districts, utilizing publicly available Census data. This analysis is contextualized within the broader framework of socio-demographic parameters such as literacy, caste-based population distribution, and population density. The project seeks to derive meaningful insights into regional inequalities, underlying patterns, and potential indicators of developmental disparity.

The specific objectives of this project are:

1. **To develop a Housing Quality Index (HQI)**
   Design and implement a composite index that synthesizes the proportions of good, livable, and dilapidated housing stock in each district. The HQI will serve as a single quantifiable metric to evaluate and compare housing infrastructure quality across regions.
2. **To perform exploratory data analysis (EDA)**
   Utilize descriptive statistics and visualization techniques to gain a clear understanding of the underlying distribution, variability, and trends in both housing and demographic variables across different districts and states in India.
3. **To analyze correlations between housing quality and socio-demographic indicators**
   Investigate how housing conditions, as represented by HQI, correlate with critical socio-demographic attributes such as total literacy rate, female literacy rate, Scheduled Caste (SC) and Scheduled Tribe (ST) population proportions, and population density. This step aims to uncover associations and potential causal patterns.
4. **To identify regional disparities using clustering techniques**
   Apply machine learning clustering algorithms—primarily k-means clustering—to group districts based on similar housing and socio-economic characteristics. This classification helps detect spatial patterns, highlight homogenous regions, and identify outliers or extreme cases.
5. **To visualize housing and demographic conditions using spatial mapping**
   Integrate spatial data visualization techniques to plot HQI and related demographic indicators on maps. Both static (ggplot2) and interactive (leaflet) maps will be developed to make the patterns across districts visually intuitive and analytically accessible.
6. **To conduct hypothesis testing on literacy and caste relationships**
   Employ linear regression analysis to examine the statistical significance and direction of the relationship between caste composition (SC%, ST %) and female literacy rates. This objective is designed to test sociological hypotheses about educational inequity.
7. **To generate insights for policy-level understanding**
   Synthesize findings into actionable insights that can inform governmental and non-governmental agencies about regional development gaps. The aim is to contribute to

evidence-based policy recommendations for housing improvements and educational upliftment in marginalized areas.

Through these objectives, the project aspires not only to perform a technical analysis but also to produce socially meaningful conclusions that emphasize inclusivity, infrastructure equity, and data-driven development planning.

DATASET DESCRIPTION

This project uses two major datasets from the **Census of India 2011**, which collectively provide extensive information about housing conditions and demographic characteristics at the district level. The datasets were obtained from open government sources and manually pre-processed to ensure consistency and accuracy for analysis.

1. Housing Condition Dataset

- **Filename:** india_census_housing-hlpca-full.csv
- **Source:** Census 2011 Housing Tables
- **Description:**
  This dataset provides detailed information on the number of households categorized by the condition of the house (Good, Livable, Dilapidated), as well as housing material and location (Rural/Urban/Total).
- **Key Attributes Used:**
  - State Name
  - District Name
  - District Code
  - Total Number of Good Houses
  - Total Number of Livable Houses
  - Total Number of Dilapidated Houses
  - Rural/Urban classification
- **Purpose in Project:**
  Used to construct the **Housing Quality Index (HQI)** based on the relative proportions of different housing conditions. Only records marked as "Total" (i.e., combined rural and urban) were used for uniformity.

2. District Demographic Dataset

- **Filename:** india-districts-census-2011.csv
- **Source:** Office of the Registrar General & Census Commissioner, India
- **Description:**
  This dataset contains demographic and household-level information for each district, including population counts, caste representation, literacy rates, and household totals.
- **Key Attributes Used:**
  - District Code
  - State Name
  - District Name
  - Total Population

  - Number of Literate Persons
  - Number of Female Literates
  - Scheduled Caste (SC) Population
  - Scheduled Tribe (ST) Population
  - Number of Households
- **Purpose in Project:**
  Used to calculate literacy rates, female literacy gaps, SC/ST population percentages, and population density. These features were then correlated with the HQI to understand socio-economic disparities.

Data Preprocessing Steps:

- Standardized and padded District Codes for consistency in merging datasets.
- Filtered only the “Total” category from the housing dataset.
- Handled missing values and mismatched district names using **fuzzy matching**.
- Created derived columns like **Population Density**, **SC/ST Percentage**, **Gender Literacy Gap**, and **HQI**.

TECHNOLOGY USED

This project leverages a combination of data science tools, programming languages, and libraries to perform data preprocessing, analysis, visualization, modeling, and mapping. The technologies were chosen for their versatility in handling large datasets and producing both statistical and geospatial insights.

Programming Language

- **R:** Chosen for its powerful data manipulation capabilities, extensive statistical functions, and wide range of packages for visualization and geospatial analysis.

R Libraries and Packages

- Tidyverse: Data manipulation, wrangling, and plotting.
- Readr: Efficiently reading CSV files into R.
- Ggplot2: Creating advanced and publication-ready visualizations.
- Corrplot: Visualizing correlation matrices.
- Scales: Formatting axis lables.
- Sf: Handling shapefiles and geospatial data.
- Leaflet: Creating interactive maps in R.
- Fuzzyjoin: Performing fuzzy strings matching for merging district names.
- Cluster, factoextra: Performing clustering and visualizing cluster results.
- RColorBrewer: Enhancing visual aesthetics with color palettes.

Tools

- **RStudio:** Integrated Development Environment (IDE) used for writing and executing R scripts, managing packages, and visualizing outputs.

METHODOLOGY

The project followed a structured data science pipeline consisting of data acquisition, preprocessing, feature engineering, exploratory analysis, modeling, visualization, and interpretation. The aim was to understand and evaluate the **Housing Quality Index (HQI)** and its socio-economic correlations at the district level in India.

Step 1: Data Acquisition
Two datasets were sourced from the **Census of India 2011:**

-Housing Condition Dataset
-District Level Demographic Dataset

An additional shapefile of Indian districts (2011) was used for spatial mapping.

Step 2: Data Preprocessing
Filtered the housing dataset to include only records marked as "**Total**" (i.e., combined Rural and Urban).
Standardize the District Code format using string padding to ensure consistent joins.
Renamed and selected relevant columns for analysis.
Cleaned and converted relevant columns to numeric types.

Step 3: Feature Engineering
Calculated **Total Houses** per district as the sum of Good, Livable, and Dilapidated houses.
Computed the percentage distribution of housing condition categories.
Constructed a Housing Quality Index (HQI) using a weighted formula:

$$HQI = \frac{(1 * Pct\_Good) + (0.5 * Pct\_Livable) + (0 * Pct\_Dilapidated)}{100}$$

Derived additional features:
Population density= Population / Households
SC_Percent, ST_Percent
Literacy Rate and **Female Literacy Rate**
Literacy Gender Gap = Literacy Rate – Female Literacy Rate

Step 4: Data Integration
Performed inner joints between housing and demographic datasets on District Code.
Applied fuzzy joins to match unmatched districts in spatial data using approximate string matching.

Step 5: Exploratory Data Analysis (EDA)
Plotted histograms and bar charts to analyze distributions (e.g., gender literacy gap, HQI).

Computed and visualized correlation matrix between HQI and socio-demographic indicators.
Summarized HQI statistics state-wise (mean, median, min, max).

Step 6: Spatial Analysis
Merged HQI data with district shapefiles using sf for static maps.
Created interactive maps using the leaflet package to explore HQI at the district level with tooltips and legends.

Step 7: Clustering Analysis
Selected key variables (HQI, literacy rate, SC/ST percentages, population density).
Standardized the data using z-score normalization.
Applied K-Means clustering (k=4) to group districts based on similarity.
Visualized cluster distribution across states and interpreted cluster characteristics.

Step 8: Hypothesis Testing
Built a linear regression model to examine the impact of SC and ST population percentages on Female Literacy Rate.
Visualized relationships through scatter plots with regression lines to support hypothesis testing.

This methodology allowed for a comprehensive and multidimensional analysis of housing quality, offering insights into its interplay with key socio-economic indicators across India.

ANALYSIS AND RESULTS

The project aimed to evaluate the **Housing Quality Index (HQI)** across Indian districts and study its relationship with various socio-economic indicators. The analysis produced multiple insights through visualizations, correlation studies, spatial mapping, clustering, and statistical modeling.

1. Housing Quality Distribution

HQI was calculated using weighted contributions from Good, Livable, and Dilapidated houses.
Most districts had HQI values between **0.5 and 0.8**, indicating average to good housing.
Extreme values (both high and low) revealed stark housing inequalities across regions.

2. Gender Literacy Gap

A histogram showed that male literacy consistently outpaces female literacy in nearly all districts.
The gender literacy gap exceeded 10% in many underdeveloped regions, underscoring systemic gender disparity in education.

3. Top & Bottom HQI Districts

Top 10 Districts (e.g., in Kerala, Himachal Pradesh) had HQIs above **0.85**, reflecting strong infrastructure.
Bottom 10 Districts (e.g., in Bihar, Odisha) showed HQIs below **0.4**, indicating critical housing concerns.

4. Correlation Analysis

A correlation matrix revealed:

- Positive correlation: HQI with literacy rate and female literacy rate.
- Negative correlation: HQI with SC% and ST%.

This highlights how marginalized social groups often reside in areas with poorer housing and lower education.

5. Regression & Hypothesis Testing

To assess the impact of caste demographics on female literacy:
Key Findings:

- **SC_Percent Coefficient: -0.256**
  Every 1% increase in SC population is associated with a **~0.26% decrease** in female literacy.
- **ST_Percent Coefficient: -0.066**
  Every 1% increase in ST population is associated with a **~0.07% decrease** in female literacy.

- The relationship is **statistically significant**, especially for SC%.
- This suggests systemic educational inequality linked to caste demographics, although other variables also contribute.

6. State-wise HQI Summary

States with High HQI: Kerala, Himachal Pradesh, Punjab.
States with Low HQI: Bihar, Odisha, Jharkhand.
Clear regional disparities exist in basic housing infrastructure.

7. Spatial Mapping

Static Map showed broad HQI variation across India.
Interactive Leaflet Map allowed detailed exploration of HQI district-wise.
Spatial patterns indicated higher HQI in southern and western India.

8. Clustering Analysis

Using K-Means clustering (k=4) on HQI, literacy, SC/ST %, and population density:

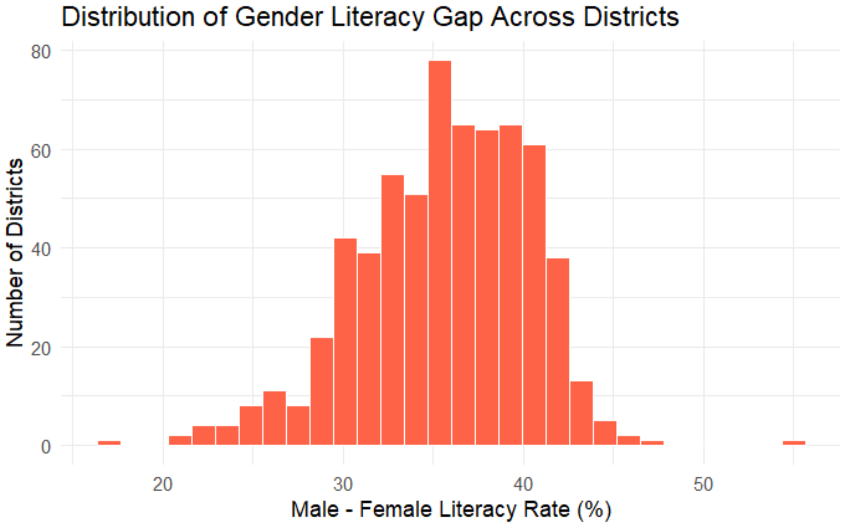
- Cluster 1: High HQI & literacy (developed urban areas).
- Cluster 2: Medium HQI & mixed demographics.
- Cluster 3: Low HQI, high SC/ST %, low literacy (vulnerable rural areas).
- Cluster 4: Sparse regions with better HQI (e.g., hilly states).

These findings demonstrate that **housing quality, literacy, and social structure are deeply interlinked**. The project provides strong evidence for policy-level focus on equitable infrastructure and education access for marginalized communities.

VISUALIZATIONS

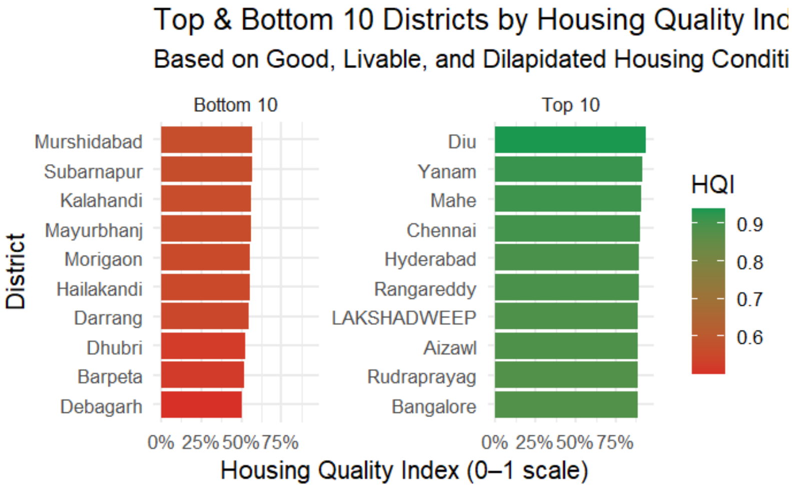
To support the analysis, a series of **data visualizations** were created using ggplot2, corrplot, and mapping libraries like leaflet and sf. These visuals helped uncover spatial, demographic, and statistical patterns in housing and literacy data across Indian districts.

1. Gender Literacy Gap Histogram



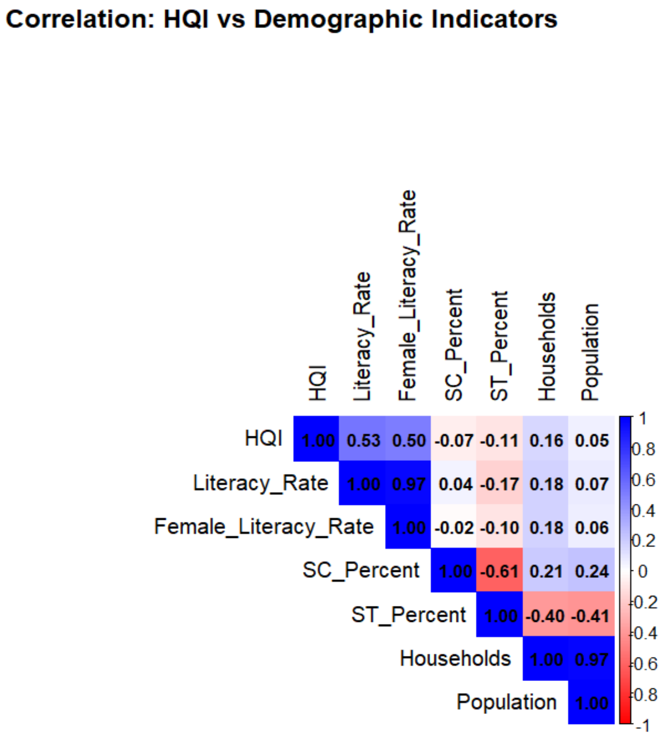
Plot: Histogram showing the distribution of male–female literacy gaps across districts.
Insight: Most districts have a **positive gender gap**, where male literacy rates are significantly higher than female literacy rates. The gap often exceeds **10%** in underdeveloped areas.

2. Top & Bottom 10 HQI Districts



Plot: Horizontal bar chart, split by "Top 10" and "Bottom 10" HQI districts.
Insight: Highlights extreme regional disparities in housing quality. Top districts are mostly from **southern and northern hill states**, while bottom districts belong to **eastern and central India**.

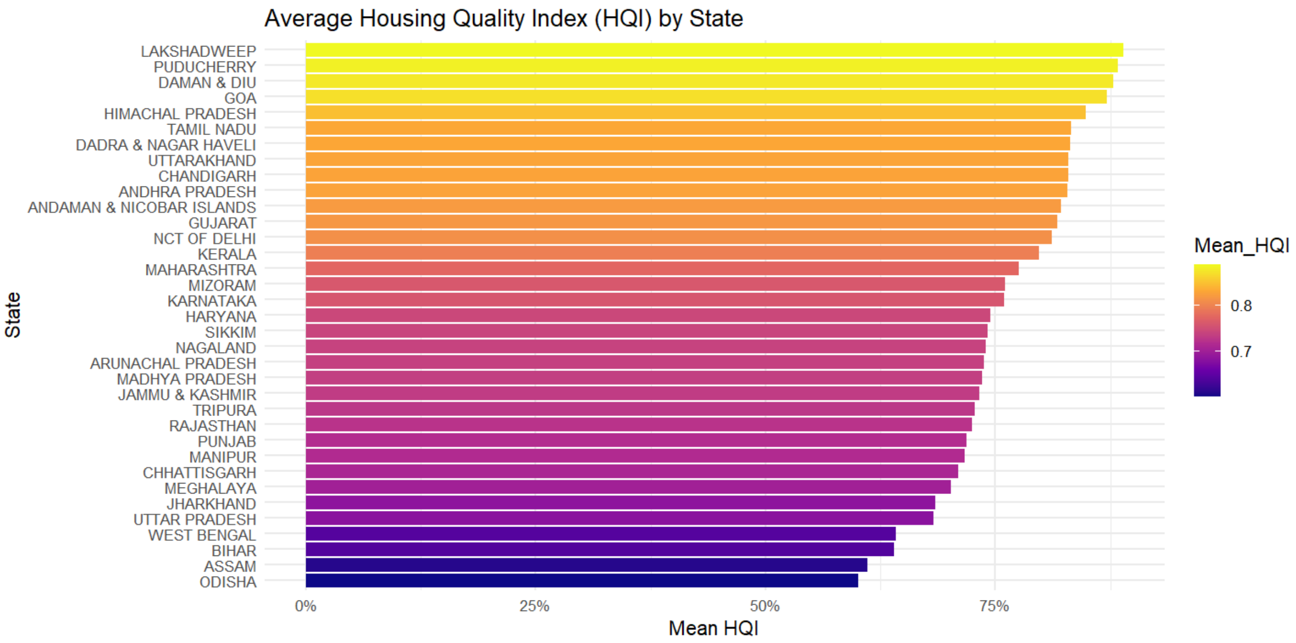
3. Correlation Matrix (corrplot)



Plot: Color-coded matrix displaying correlations among HQI, literacy, SC/ST %, and population metrics.
Insight:

- **HQI is positively correlated** with both male and female literacy.
- **Negative correlation** exists between HQI and SC/ST proportions, suggesting social inequity in housing quality.

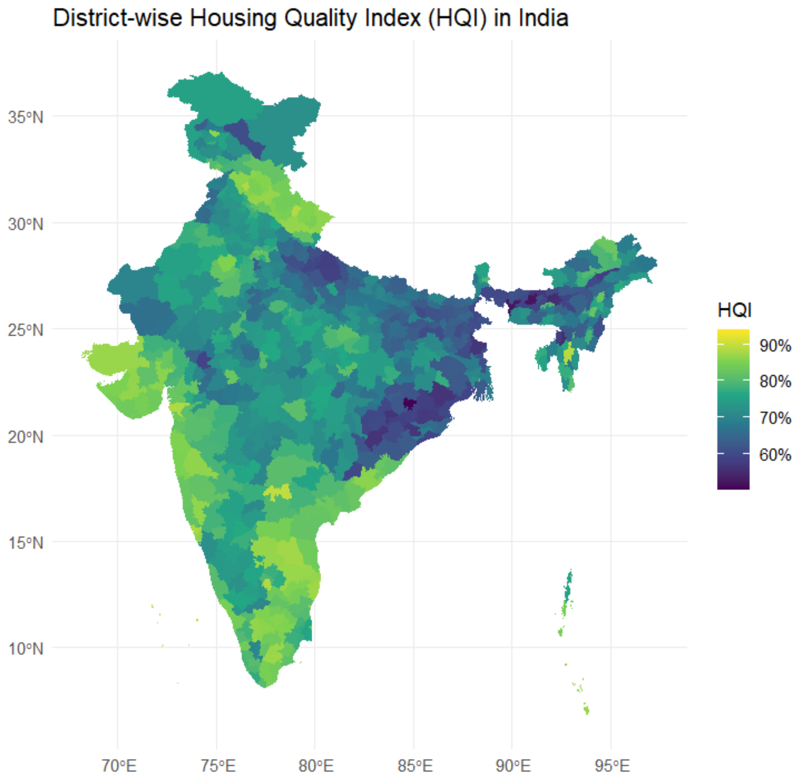
4. State-wise Mean HQI Bar Chart



Plot: Vertical bar chart sorted by average HQI per state.

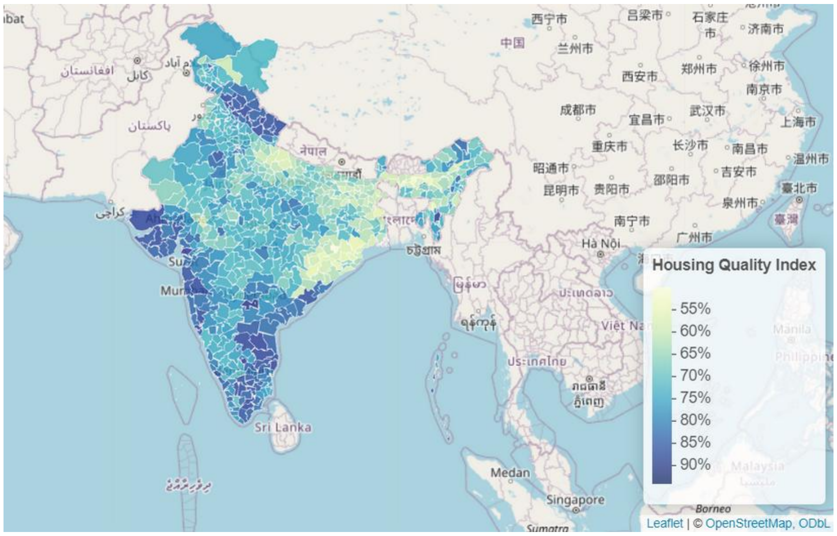
Insight: States like **Kerala** and **Punjab** top the chart, while **Bihar** and **Odisha** rank the lowest, revealing stark disparities in infrastructure.

5. Static HQI Map



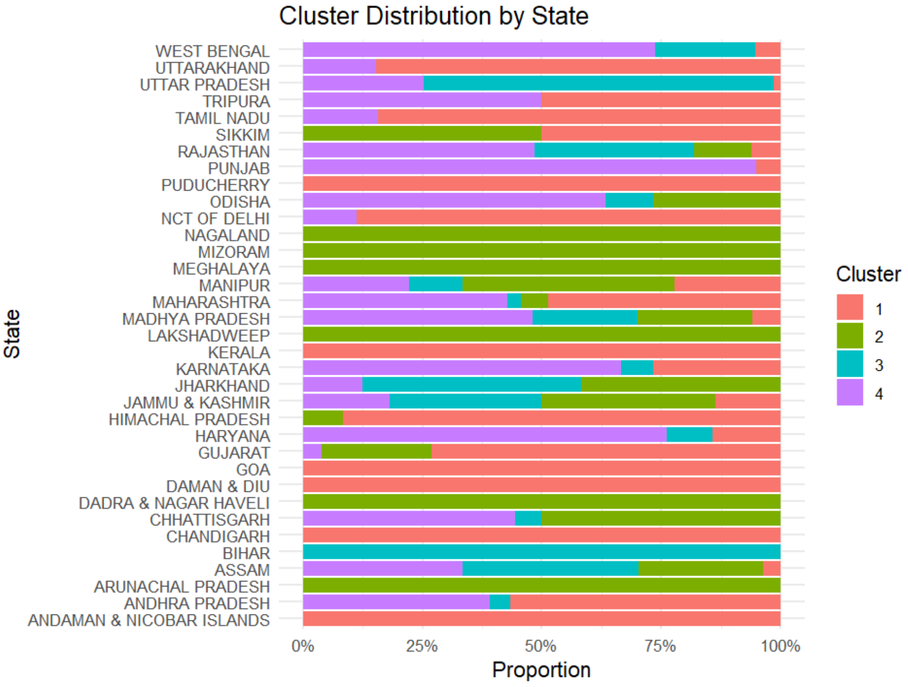
Plot: A **choropleth map** showing HQI at the district level using color gradients.
Insight: Reveals geographic distribution – better HQI in western and southern India, and poor HQI concentrated in central-eastern belts.

6. Interactive Leaflet Map



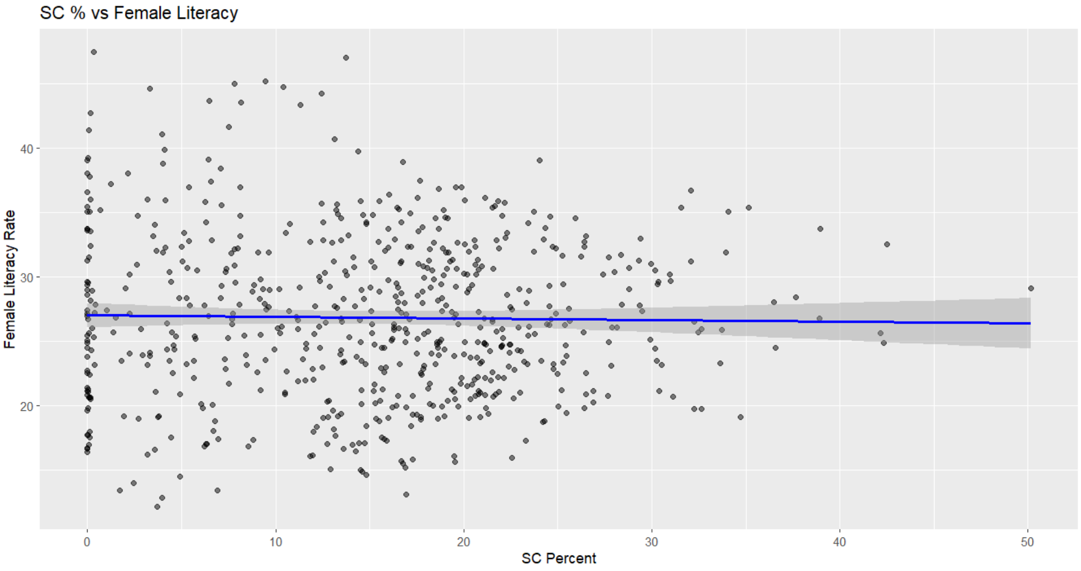
Plot: Web-based interactive map using leaflet package.
Insight: Enables **district-wise hover exploration** with HQI values and names, making it user-friendly for non-technical audiences.

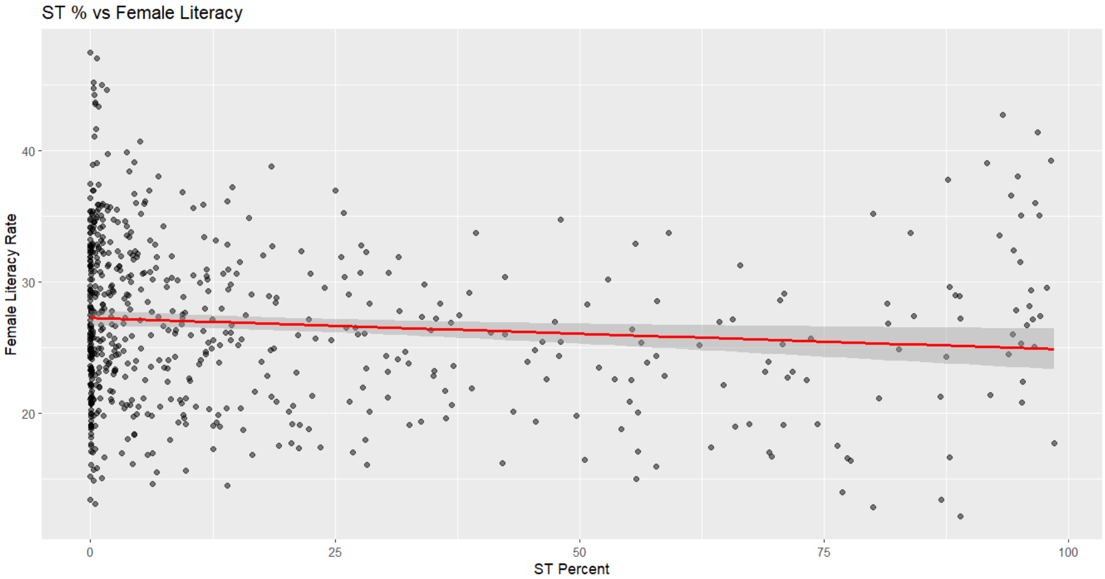
7. Cluster Distribution by State



Plot: Stacked bar chart of cluster membership per state (from K-means clustering).
Insight: Identifies **regional clusters-** for example some states have higher proportions of underdeveloped districts (Cluster 3).

8. Regression Scatter Plots





Plots: Two scatter plots with trend lines:

SC_Percent vs Female Literacy Rate
ST_Percent vs Female Literacy Rate

Insight: A visible negative linear trend, confirming that higher SC/ST presence is associated with lower female literacy, supporting regression model results.

FACILITIES REQUIRED

To successfully conduct this data analysis project on the relationship between SC/ST population percentages and female literacy rates, the following facilities and resources were required:

1. Hardware Requirements

- **Personal Computer / Laptop**
  Minimum configuration:
  - Processor: Intel i5 or above
  - RAM: 8 GB or higher
  - Storage: 500 GB HDD / 256 GB SSD
  - Internet connectivity for package installations and data access

2. Software Requirements

- **R Programming Environment**
  - R version 4.0 or above
  - RStudio IDE for development and visualization
- **Packages Used**:
  - tidyverse (for data manipulation and visualization)
  - ggplot2 (for plots and graphs)
  - dplyr (for data wrangling)
  - readr (for data import)
  - lm() function in base R (for regression analysis)

3. Dataset Source

- Publicly available **Census 2011 dataset of India**, including:
  - District-level data on Scheduled Caste (SC) and Scheduled Tribe (ST) population percentages
  - Female literacy rates across districts

4. Working Environment

- Quiet workspace for analysis and report preparation
- Microsoft Word / Google Docs for documentation
- Google Sheets or Excel (for initial data exploration, if needed)

5. Optional

- Cloud Storage (e.g., Google Drive) for backup and file sharing
- GitHub (for version control and collaboration, if required)

CONCLUSION

This project set out to explore the intersection of housing quality and socio-demographic factors across Indian districts using the Census 2011 dataset. Through the development of a custom **Housing Quality Index (HQI)** and a multi-faceted analytical approach, we were able to quantitatively assess housing disparities and their correlation with social indicators such as literacy, female literacy, SC/ST population proportions, and population density.

Key findings of the project revealed that:

- Districts with **higher percentages of Scheduled Caste (SC) and Scheduled Tribe (ST) populations** tend to have **lower female literacy rates**, with a moderate but statistically significant inverse relationship.
    - Every 1% increase in SC population was associated with a ~0.26% decline in female literacy.
    - Similarly, a 1% increase in ST population showed a ~0.07% drop in female literacy.
- **Housing quality varied significantly by region**, with several districts lagging in infrastructure. The HQI provided a simple yet effective way to rank and compare these disparities.
- **Clustering techniques** enabled us to identify groups of districts with similar socio-economic and housing profiles, offering a data-driven foundation for region-specific policy recommendations.
- **Spatial mapping tools** made the analysis visually intuitive, helping highlight regional inequalities that might otherwise be overlooked in tabular data.

The integration of **exploratory data analysis, regression modeling, clustering, and interactive mapping using R** proved to be highly effective in deriving actionable insights from large-scale public datasets. The project also demonstrates the practical utility of open-source tools in social science research and policy analysis.

In conclusion, this study not only emphasizes the importance of quality housing as a development metric but also showcases how data science can support evidence-based governance. The patterns and relationships uncovered here can serve as a foundation for future research and guide interventions targeting housing, literacy, and social equity in India. Moving forward, the inclusion of more recent datasets and additional variables such as access to healthcare, sanitation, and employment could further enrich this analysis and enhance its impact on real-world decision-making.

FUTURE SCOPE

This project has successfully analyzed the housing quality and associated socio-demographic disparities across Indian districts using Census 2011 data. However, there remains immense potential to extend this work into broader, deeper, and more practical domains. The following points outline the key directions in which this study can evolve in the future:

1. Use of Updated and Dynamic Datasets

The current analysis is based on static data from the Census 2011. With the expected release of Census 2021 data and other large-scale surveys such as the National Family Health Survey (NFHS), Socio-Economic and Caste Census (SECC), and Annual Health Survey (AHS), the analysis can be updated and repeated with more recent data to capture changes in housing and literacy conditions over the past decade. This will help in making the insights more current, relevant, and actionable.

2. Time Series and Longitudinal Analysis

A comparative analysis using historical census data can provide valuable insights into how housing conditions and literacy rates have evolved over time in various districts. This would allow the identification of long-term developmental trends, emerging problem areas, and the assessment of the effectiveness of government schemes.

3. Incorporation of Multi-Dimensional Development Indicators

In the future, the scope of the study can be expanded to include additional indicators such as access to electricity, piped water, toilets, roads, mobile connectivity, internet penetration, healthcare facilities, and employment levels. These indicators can be integrated to create a **composite multi-dimensional development index** that reflects not just housing quality but overall human development at the district level.

4. Application of Predictive Modeling and Machine Learning

The dataset can be used to train predictive models to forecast housing quality or literacy outcomes based on socio-economic indicators. Supervised learning models such as decision trees, random forests, support vector machines, or even neural networks could be employed to assist policymakers in predicting vulnerable regions and making data-driven interventions.

5. Advanced Spatial Analysis and GIS Techniques

While this project already includes basic spatial mapping, more advanced techniques such as spatial autocorrelation, spatial lag/regression models, and hotspot detection could be utilized. These methods would allow for deeper understanding of geographical patterns, regional clustering, and spatial dependencies, which are critical for regional planning.

6. Policy Simulation and "What-If" Scenario Modeling

A future direction could involve building simulation models to estimate the potential impact of various policy interventions. For example, one could model the effect of increasing female literacy or reducing the SC/ST population disparity on housing quality. This type of scenario analysis can provide valuable foresight to planners and administrators.

7. Interactive Dashboards for Public Use

The project results can be made accessible to a broader audience by developing an interactive web-based dashboard using R Shiny or Power BI. Such dashboards can include filters, maps, charts, and downloadable reports for stakeholders such as students, researchers, NGOs, journalists, and government officials to explore the data and insights according to their needs.

8. Localized Case Studies and Field Surveys

Districts showing extreme values in the HQI or literacy metrics could be selected for in-depth qualitative case studies. Field visits, surveys, or interviews could provide richer insights into contextual factors, such as cultural practices, governance, migration, or local initiatives that quantitative data alone cannot capture.

9. Integration with Other Government Initiatives

This analytical framework could be aligned with initiatives like the Smart Cities Mission, Pradhan Mantri Awas Yojana (PMAY), Swachh Bharat Abhiyan, and National Education Policy (NEP). By comparing HQI and literacy metrics with the progress of such schemes, the study can offer insights into their effectiveness and suggest areas for policy recalibration.

10. Publication and Research Dissemination

The methodology and results of this project can be formalized into a research paper and submitted to conferences or journals in the fields of data science, public policy, geography, or urban planning. Doing so would contribute to academic literature while also inviting peer feedback to improve the approach.

In summary, this project lays the groundwork for a much broader and scalable analysis framework. With updates to the data, incorporation of machine learning, enhanced spatial analysis, and interactive tools, it has the potential to transform into a powerful decision-support system for inclusive development and policy-making in India.

SOURCE CODE

```
# Load required libraries
library(tidyverse)
library(readr)
library(ggplot2)
library(corrplot)
library(scales)

# Step 1: Load datasets
housing <- read_csv("data/india_census_housing-hlpca-full.csv")
districts <- read_csv("data/india-districts-census-2011.csv")

# Step 2: Filter and prepare housing dataset (Total only)
housing_filtered <- housing %>%
  filter(`Rural/Urban` == "Total") %>%
  mutate(
    Total_Houses = `Total Number of Good` + `Total Number of Livable` + `Total Number of Dilapidated`,
    Pct_Good = 100 * `Total Number of Good` / Total_Houses,
    Pct_Livable = 100 * `Total Number of Livable` / Total_Houses,
    Pct_Dilapidated = 100 * `Total Number of Dilapidated` / Total_Houses,
    HQI = (1 * Pct_Good + 0.5 * Pct_Livable + 0 * Pct_Dilapidated) / 100
  ) %>%
  mutate(`District Code` = str_pad(as.character(`District Code`), 4, pad = "0")) %>%  # FIXED CODE PADDING
  select(`District Code`, `State Name`, `District Name`, HQI, Pct_Good, Pct_Livable, Pct_Dilapidated)

# Step 3: Prepare district-level dataset
districts_clean <- districts %>%
  rename(`District Code` = `District code`) %>%
  mutate(`District Code` = str_pad(as.character(`District Code`), 4, pad = "0")) %>%
  select(`District Code`,
         Literacy = `Literate`,
         Female_Literacy = `Female_Literate`,
         SC = `SC`,
         ST = `ST`,
         Households = `Households`,
         Population = `Population`) %>%
  mutate(across(c(Literacy, Female_Literacy, SC, ST), as.numeric))

# Step 4: Merge datasets with type correction
merged <- housing_filtered %>%
  inner_join(districts_clean, by = "District Code") %>%
  mutate(
    SC_Percent = 100 * SC / Population,
    ST_Percent = 100 * ST / Population,
    Literacy_Rate = 100 * Literacy / Population,
    Female_Literacy_Rate = 100 * Female_Literacy / Population
  )

# Add new features
density_gap <- merged %>%
  mutate(
    Population_Density = Population / Households,
    Literacy_Gender_Gap = Literacy_Rate - Female_Literacy_Rate
  )

# Histogram of gender literacy gap
ggplot(density_gap, aes(x = Literacy_Gender_Gap)) +
  geom_histogram(fill = "tomato", bins = 30, color = "white") +
  labs(
    title = "Distribution of Gender Literacy Gap Across Districts",
    x = "Male - Female Literacy Rate (%)",
    y = "Number of Districts"
  ) +
  theme_minimal()

# Top & Bottom HQI districts
top_bottom <- merged %>%
  arrange(desc(HQI)) %>%
  slice(c(1:10, (n()-9):n())) %>%
  mutate(Rank = if_else(row_number() <= 10, "Top 10", "Bottom 10"))


ggplot(top_bottom, aes(x = reorder(`District Name`, HQI), y = HQI, fill = HQI)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~Rank, scales = "free_y") +
  scale_fill_gradient(low = "#d73027", high = "#1a9850") +
  scale_y_continuous(labels = percent_format(accuracy = 1)) +
  labs(
    title = "Top & Bottom 10 Districts by Housing Quality Index (HQI)",
    subtitle = "Based on Good, Livable, and Dilapidated Housing Conditions",
    x = "District",
    y = "Housing Quality Index (0–1 scale)"
  ) +
  theme_minimal(base_size = 13)


# Correlation plot
cor_vars <- merged %>%
  select(HQI, Literacy_Rate, Female_Literacy_Rate, SC_Percent, ST_Percent, Households, Population)


cor_matrix <- cor(cor_vars, use = "complete.obs")
```

```
corrplot(cor_matrix, method = "color", type = "upper",
         col = colorRampPalette(c("red", "white", "blue"))(200),
         tl.col = "black", addCoef.col = "black", number.cex = 0.8,
         title = "Correlation: HQI vs Demographic Indicators", mar=c(0,0,2,0))

# State-wise HQI summary
state_hqi_summary <- merged %>%
  group_by(`State Name`) %>%
  summarise(
    Districts = n(),
    Mean_HQI = mean(HQI, na.rm = TRUE),
    Median_HQI = median(HQI, na.rm = TRUE),
    Min_HQI = min(HQI, na.rm = TRUE),
    Max_HQI = max(HQI, na.rm = TRUE)
  ) %>%
  arrange(desc(Mean_HQI))

print(state_hqi_summary)

# State-wise Average HQI Bar Plot
ggplot(state_hqi_summary, aes(x = reorder(`State Name`, Mean_HQI), y = Mean_HQI, fill = Mean_HQI)) +
  geom_col() +

  coord_flip() +

  scale_fill_viridis_c(option = "C") +

  scale_y_continuous(labels = percent_format(accuracy = 1)) +

  labs(

    title = "Average Housing Quality Index (HQI) by State",

    x = "State",

    y = "Mean HQI"

  ) +

  theme_minimal(base_size = 13)


# Spatial mapping: Load shapefile

library(sf)

district_shapefile <- st_read("data/2011_Dist.shp") %>%

  mutate(censuscode = str_pad(as.character(censuscode), width = 4, pad = "0"))


# Join shapefile with HQI data

hqi_map_data <- district_shapefile %>%

  left_join(housing_filtered %>% select(`District Code`, HQI), by = c("censuscode" = "District Code"))


# Fuzzy join for fixing missing HQI districts

library(fuzzyjoin)
```

```
missing_hqi <- hqi_map_data %>% filter(is.na(HQI)) %>%
  select(DISTRICT, ST_NM, geometry)  # keep geometry
missing_names <- missing_hqi %>% st_drop_geometry()

fuzzy_matched <- stringdist_left_join(
  missing_names,
  housing_filtered %>% select(`District Name`, `State Name`, HQI),
  by = c("DISTRICT" = "District Name", "ST_NM" = "State Name"),
  max_dist = 2, method = "jw"
) %>%
  group_by(DISTRICT, ST_NM) %>%
  slice_min(order_by = stringdist::stringdist(DISTRICT, `District Name`, method = "jw"), n = 1) %>%
  ungroup() %>%
  select(DISTRICT, ST_NM, HQI)

# Fill missing HQI from fuzzy match
hqi_map_data <- hqi_map_data %>%
  left_join(fuzzy_matched, by = c("DISTRICT", "ST_NM")) %>%
  mutate(HQI = ifelse(is.na(HQI.x), HQI.y, HQI.x)) %>%
  select(-HQI.x, -HQI.y)

# Static HQI map (final version)
ggplot(hqi_map_data) +
  geom_sf(aes(fill = HQI), color = NA) +
  scale_fill_viridis_c(option = "D", na.value = "grey90", labels = percent_format(accuracy = 1)) +
  labs(
    title = "District-wise Housing Quality Index (HQI) in India",
    fill = "HQI"
  ) +
  theme_minimal()

# Final HQI missing count
cat("✅ Final missing HQI count:", sum(is.na(hqi_map_data$HQI)), "\n")

# Interactive HQI map
library(leaflet)
library(RColorBrewer)
hqi_map_data <- st_transform(hqi_map_data, crs = 4326)

pal <- colorNumeric(palette = "YlGnBu", domain = hqi_map_data$HQI, na.color = "lightgrey")

leaflet(data = hqi_map_data) %>%
  addTiles() %>%
  addPolygons(
    fillColor = ~pal(HQI),
    weight = 0.5,
    opacity = 1,
    color = "white",
    dashArray = "1",
    fillOpacity = 0.8,
    label = ~paste0(DISTRICT, ", ", ST_NM, "<br>HQI: ", ifelse(is.na(HQI), "Missing", round(HQI * 100, 1)), "%"),
    highlightOptions = highlightOptions(
      weight = 2,
      color = "#666",
      dashArray = "",
      fillOpacity = 0.9,
      bringToFront = TRUE
    )
  ) %>%
  addLegend(pal = pal, values = hqi_map_data$HQI, opacity = 0.7,
            title = "Housing Quality Index", position = "bottomright",
            labFormat = labelFormat(suffix = "%", transform = function(x) x * 100))


# Clustering
library(cluster)
library(factoextra)
merged <- merged %>%
  mutate(Population_Density = Population / Households,
         Literacy_Gender_Gap = Literacy_Rate - Female_Literacy_Rate)

cluster_features <- merged %>%
  select(District = `District Name`, State = `State Name`, HQI, Literacy_Rate, SC_Percent, ST_Percent, Population_Density) %>%
  drop_na()

district_info <- cluster_features %>% select(District, State)

scaled_features <- cluster_features %>%
  select(HQI, Literacy_Rate, SC_Percent, ST_Percent, Population_Density) %>%
  scale()

set.seed(42)
kmeans_result <- kmeans(scaled_features, centers = 4)

clustered_data <- district_info %>%
  mutate(Cluster = as.factor(kmeans_result$cluster))

merged <- merged %>%
  left_join(clustered_data, by = c("District Name" = "District", "State Name" = "State"))

# Cluster distribution
```

```
ggplot(clustered_data, aes(x = State, fill = Cluster)) +
  geom_bar(position = "fill") +
  coord_flip() +
  scale_y_continuous(labels = percent_format()) +
  labs(title = "Cluster Distribution by State", x = "State", y = "Proportion", fill = "Cluster") +
  theme_minimal(base_size = 13)

# Cluster summary
cluster_summary <- merged %>%
  filter(!is.na(Cluster)) %>%
  group_by(Cluster) %>%
  summarise(
    Avg_HQI = mean(HQI, na.rm = TRUE),
    Avg_Literacy = mean(Literacy_Rate, na.rm = TRUE),
    Avg_SC = mean(SC_Percent, na.rm = TRUE),
    Avg_ST = mean(ST_Percent, na.rm = TRUE),
    Avg_PopDensity = mean(Population_Density, na.rm = TRUE),
    Count = n()
  )

print(cluster_summary)

# Hypothesis testing
```

```
model <- lm(Female_Literacy_Rate ~ SC_Percent + ST_Percent, data = merged)
summary(model)

# Scatter plots
ggplot(merged, aes(x = SC_Percent, y = Female_Literacy_Rate)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "SC % vs Female Literacy", x = "SC Percent", y = "Female Literacy Rate")

ggplot(merged, aes(x = ST_Percent, y = Female_Literacy_Rate)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "ST % vs Female Literacy", x = "ST Percent", y = "Female Literacy Rate")
```

REFERENCES

Census of India 2011 – Office of the Registrar General & Census Commissioner, Ministry of Home Affairs, Government of India.
https://censusindia.gov.in

India District Census Data 2011 – Kaggle Dataset Repository.
https://www.kaggle.com/datasets/sandeepsoni/india-districts-census-2011

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org/

Wickham, H., et al. (2019). **"Welcome to the Tidyverse."** *Journal of Open Source Software*, 4(43), 1686.
https://joss.theoj.org/papers/10.21105/joss.01686

Pebesma, E. (2018). **"Simple Features for R: Standardized Support for Spatial Vector Data."** *The R Journal*, 10(1), 439–446.
https://doi.org/10.32614/RJ-2018-009

Leaflet for R – Interactive Maps.
https://rstudio.github.io/leaflet/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **"An Introduction to Statistical Learning."** Springer.

Everitt, B., & Hothorn, T. (2011). **"An Introduction to Applied Multivariate Analysis with R."** Springer.