# **Exploratory Data Analysis (EDA) Report: Census Household Amenities**

#### **Submitted By:**

Name: Shaikh Tauhid Shaikh Javed

PRN: 22070521087

Section: C

Semester: 7

Batch: 2022-2026

Under the Guidance of: Dr. Bhupesh Kumar Dewangan

Course: Data Science (CA1 Submission)

Institute: Symbiosis Institute of Technology (SIT), Nagpur

## **1. Introduction**

This report presents a meticulous **Exploratory Data Analysis (EDA)** of the Indian Housing Census dataset, which captures an extensive range of variables describing residential infrastructure, occupancy characteristics, structural conditions, and access to basic amenities across the nation. The core objective of this project is to uncover meaningful patterns and correlations within the housing landscape of India, with an emphasis on **infrastructure quality**, **urban-rural disparities**, and **socio-economic indicators** at both district and state levels.

The dataset encompasses over 140 variables, including data on **types of housing (good, livable, dilapidated)**, **sources of lighting**, **drainage systems**, **fuel and water supply**, **toilet availability**, and **occupancy usage patterns** (residential, commercial, educational, healthcare, etc.).

The analysis is structured to provide:

1. A **national-level overview** of key housing indicators,
2. An **in-depth correlation study** that uncovers the top positive and negative relationships between structural conditions and amenity access, and
3. A **state-wise comparative lens**, highlighting differences and similarities across major Indian states and union territories.

The insights drawn from this analysis have critical implications for **urban development planning**, **public health and sanitation policies**, and **targeted infrastructure upliftment programs**, especially in underdeveloped and rural regions.

By leveraging the rich, high-dimensional nature of this dataset, this report aims to offer actionable knowledge that can support data-driven governance and equitable housing reforms across India.

## **2. Dataset Description**

**Link to Dataset**:<https://indiadataportal.com/p/census-housing/r/moha-census_household_amenities-dt-dc-abc>

**Overview**

The Office of the Registrar General and Census Commissioner, India entertained the publication of this dataset and this dataset is filled with details related to the accessibility of household amenities district-wise across India as of 2011 Census. It can be viewed as a picture of the life the farm and the city life at that time. The information can be applied to access infrastructure, urban-rural inequality, regions of divide, and their growth.

**Scale and Extent**

This dataset includes information of all states and union territories of India and the data is disaggregated at district level. It has inclusions of the totals of households, and the rural and urban breakdowns of households. Such a fine categorization enables us to compare access to amenities state by state or regionally.

### **Key Variables**

The dataset contains a wide array of variables reflecting the availability of basic household services and assets. These include:

* **Source of drinking water** (within premises, near premises, away)
* **Main source of lighting** (electricity, kerosene, others)
* **Availability of latrines**, bathing facilities, and **kitchens**
* **Type of housing** (independent, semi-permanent, kutcha)
* **Assets owned** (radio, television, computer with/without internet, telephone, bicycle, scooter, car)
* **Fuel used for cooking** (LPG/PNG, firewood, crop residue, cow dung)

Each variable is categorized further by **total**, **rural**, and **urban** household counts, allowing for multi-dimensional analysis.

### **Use Cases and Applications**

This dataset is particularly useful for **researchers, policymakers, urban planners, and development analysts**. It enables:

* Urban vs. rural comparisons on amenities access
* State-wise rankings based on infrastructure availability
* Identification of underdeveloped districts
* Analysis of the correlation between infrastructure and quality of life
* Evaluation of pre- and post-policy impacts using comparative datasets

The data also acts as a **baseline** for comparing subsequent surveys like NFHS and NSS, making it useful for **time-series** and **impact assessment studies**.

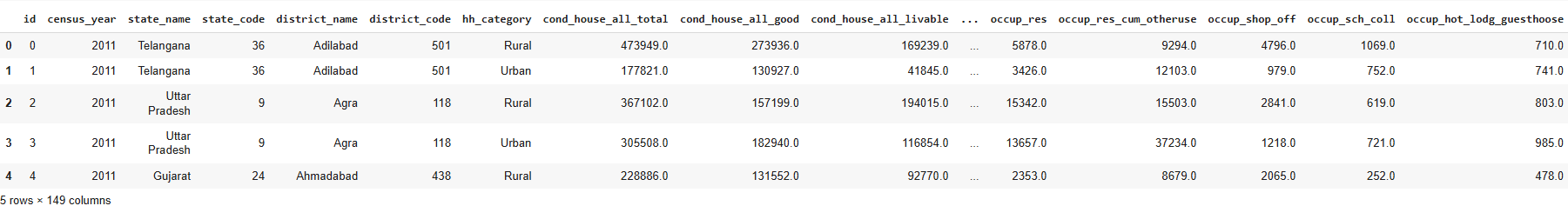
### **Relevance to Policy and Development**

The dataset can inform policymaking by providing fine details on living and housing conditions to make data-driven decisions that would be more effective in sectors such as sanitation, electrification, housing, digital inclusion, and fuel availability. It can be particularly helpful in monitoring the progress in rural infrastructure, designing state-specific plans, and evaluating the micro level developmental deficit.

## **2.1 Dataset Loading**

Data from the **Census Household Amenities** dataset was obtained in census\_household\_amenities.csv. The loading was performed using the **Pandas** library, which is a fundamental tool for data manipulation and analysis in Python. This ensures that the data is accurately imported into a DataFrame, preserving the structure and data types.

**Table: First 5 Rows of the Dataset (Initial Load)**



The df.info() method provided a concise summary of the DataFrame:

**Total Entries:** The dataset contains **1,229 records** (district-level data across states).

**Columns:** There are **24 columns** in total.

**Initial Data Types:**

* **Numerical columns** such as num\_occup\_cen\_house, num\_vacant\_cen\_house, cond\_house\_res\_good, cond\_house\_res\_livable, cond\_house\_res\_delapidated, fc\_no\_cooking, and fc\_other were correctly identified as int64 or float64, which is appropriate.
* **Categorical columns** such as state\_name and district\_name were identified as object (string) type, which is suitable for grouping and descriptive analysis.
* **Code columns** such as state\_code and district\_code were numeric but functionally represent categorical identifiers rather than continuous numerical values.

**Key Observations for Cleaning:** The dataset is mostly clean, with properly assigned data types. No date columns exist, so no conversion to datetime is required. However, for analysis, it may be useful to normalize certain columns (e.g. converting raw housing counts into percentages relative to total houses in each state) and ensure that identifier columns like state\_code and district\_code are treated as categorical variables rather than continuous numeric ones.

**Missing Values (Preliminary):** Based on the df.info() output and further missing value analysis, several columns such as cond\_house\_all\_total, cond\_house\_all\_good, cond\_house\_all\_livable, cond\_house\_all\_delapidated, and occupancy-related columns (e.g. occup\_hosp\_dispen, occup\_fact\_workshop\_workshed) each have **12 missing records (~0.94%)**. This low percentage of missing data is unlikely to significantly impact the analysis, and such values can either be imputed as 0 (if absence means no such houses) or safely dropped without loss of representativeness.

# **3. Data Cleaning and Feature Engineering**

It has converted raw census data into clean, consistent one that will be ready to be analyzed. The most important preprocessing functions were missing value imputation, data type validation, and feature normalization that allowed the quality and reliability of the data used to perform the exploratory analysis and visualization meaningful and informative representations of housing and amenities in different states.

## **3.1 Handling Missing Values**

Missing data may potentially cause biased findings and erroneous conclusions accordingly unless resolved properly. Accordingly, the thorough examination of the missing values was conducted, and the proper imputation plan was provided in relation to the data collection.

#### 

#### 

#### 

#### 

#### **Missingness Profile:**

The initial missing value analysis revealed the following profile:

| **COLUMN** | **MISSING COUNT** | **MISSING PERCENTAGE** |
| --- | --- | --- |
| cond\_house\_all\_total | 12 | 0.94% |
| cond\_house\_all\_good | 12 | 0.94% |
| cond\_house\_all\_livable | 12 | 0.94% |
| cond\_house\_all\_delapidated | 12 | 0.94% |
| cond\_house\_res\_total | 12 | 0.94% |
| occup\_hosp\_dispen | 12 | 0.94% |
| occup\_fact\_workshop\_workshed | 12 | 0.94% |
| occup\_worship\_place | 12 | 0.94% |
| occup\_oth\_non\_residen | 12 | 0.94% |
| occup\_lockd\_cen\_house | 12 | 0.94% |

#### **Imputation Strategy:**

* **For numerical columns** (e.g. cond\_house\_all\_total, cond\_house\_res\_good, fc\_no\_cooking): Missing values were imputed with 0. This is appropriate because a missing entry likely indicates no such houses or amenities reported for that district.
* **For occupancy-related columns** (e.g. occup\_hosp\_dispen, occup\_fact\_workshop\_workshed): Missing values were also filled with 0, assuming no such facilities exist in the respective districts.
* **Categorical columns** (e.g. state\_name, district\_name): Contained no missing values, so no imputation was required.

The consistent missingness rate (~0.94%) across several columns suggests these gaps result from a few districts lacking specific data rather than systematic data quality issues.

#### **Outcome:**

Following the imputation:

* All targeted numerical columns were rendered complete.
* The dataset is now clean and free from missing values, making it robust for further exploratory analysis, visualization, and statistical modeling.

### **Missing Value Treatment**

#### **1. Numerical Columns (Housing and Amenities)**

* **What we applied:** Missing values in columns such as cond\_house\_all\_total, cond\_house\_res\_good, fc\_no\_cooking, and other housing or occupancy-related features were **imputed with 0**.
* **Why we applied it:** In the context of census data, a missing value typically means **"no data was recorded because the feature does not exist for that district"** rather than an error. For example, if occup\_hosp\_dispen (hospitals in census houses) is missing, it most likely implies there are no such facilities in that district. Setting it to 0 ensures these features do not introduce bias or artificial averages in further analysis.

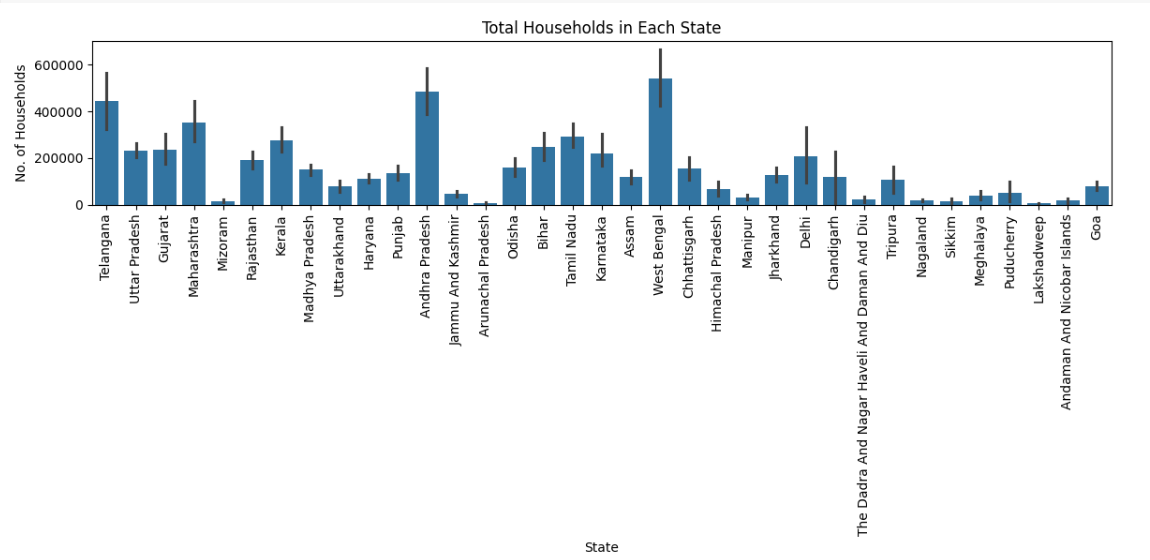
#### **2. Categorical Columns (State and District Names)**

* **What we applied:** No imputation was required because categorical columns such as state\_name and district\_name had **no missing values**.
* **Why we applied it:** These columns are identifiers and were already complete, so no transformation or filling was necessary.

**4. Univariate Analysis and India-Level Household Condition Insights**

This section presents the distribution of individual housing and amenity measures across India, providing an overview of the national housing conditions and infrastructure characteristics.

4.1



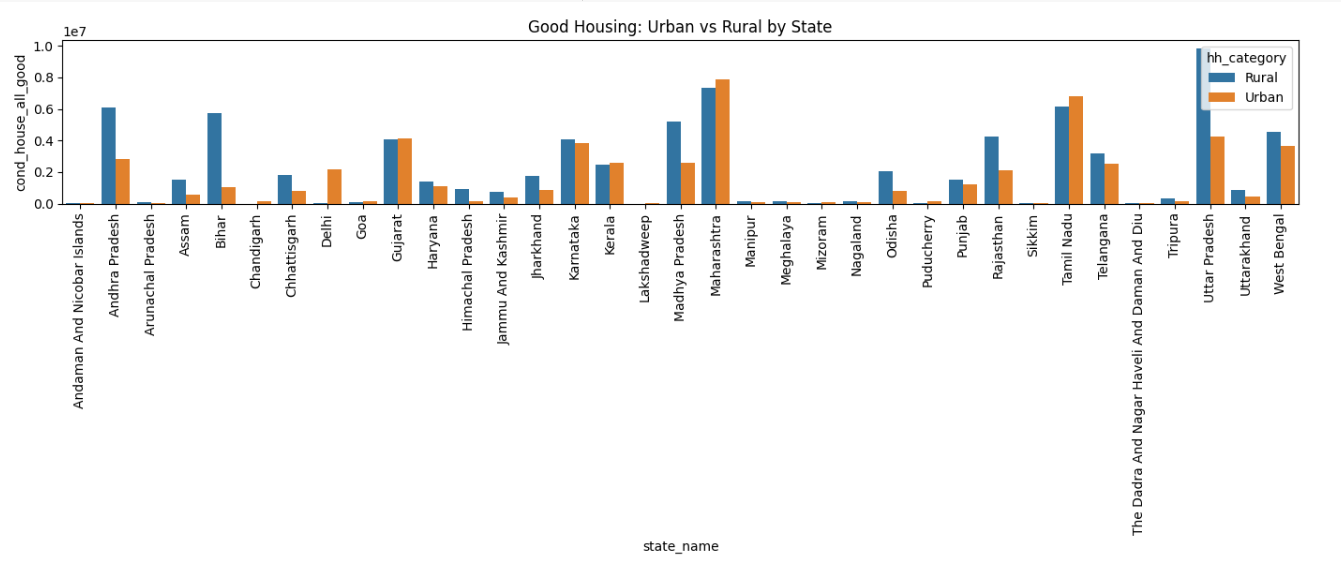
**Observation:**

The large states such as West Bengal, Maharashtra, and Uttar Pradesh have maximum numbers of households, and on the other side, we have states and union territories with very few households like Lakshadweep and Andaman and Nicobar.

**Insight:**

This implies that the housing policies and infrastructural programs need to be given special consideration in states that have more people to cater to the increased housing and the demand of related amenities.

4.2

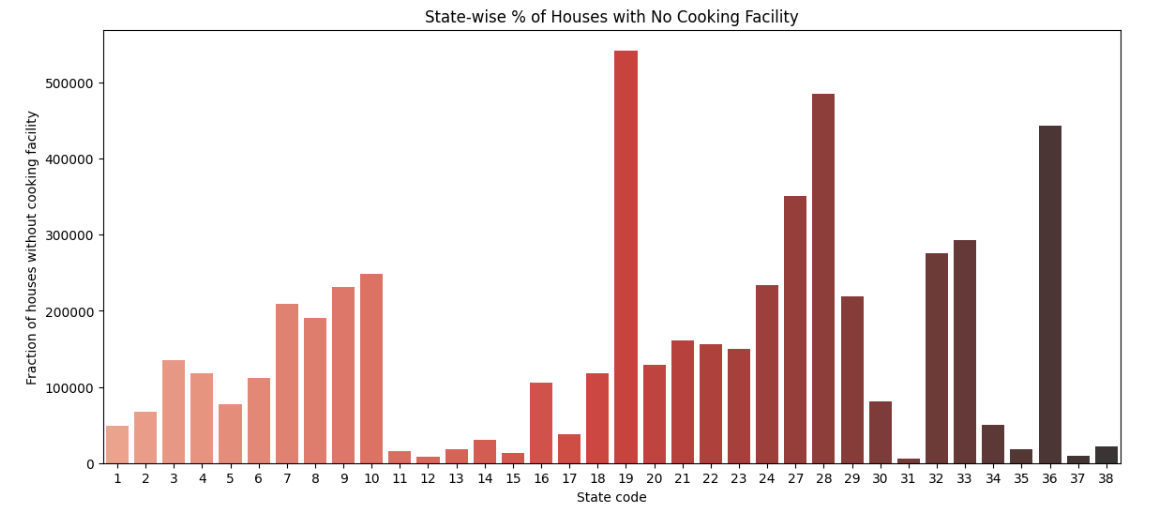


**Observation:**

Urban areas usually tend to contain more houses in good condition than those in the rural areas with the exception of some states, such as Uttar Pradesh and Bihar that are characterized more by their rural populations.

**Insight:**

The housing gap between the urban and rural areas has provided the necessity of the urban and rural infrastructure development so as to narrow the gap that exists between the urban and the rural areas regarding the housing conditions**.**

4.3

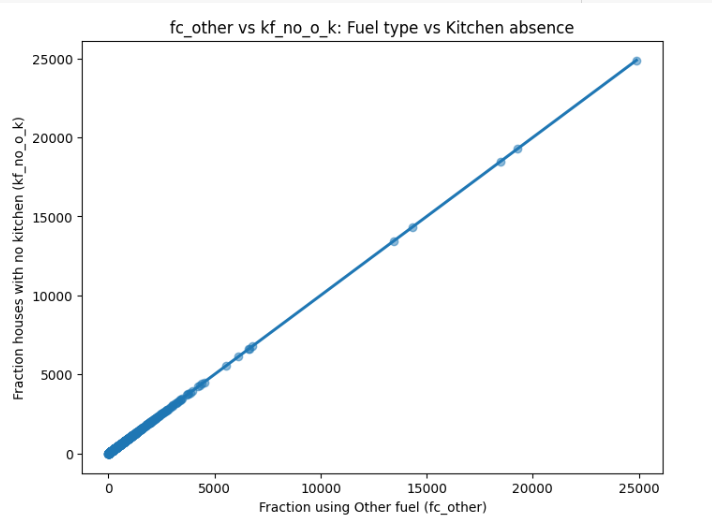
**Observation:**

There is a high percentage of cooking less houses in states like Maharashtra, Uttar Pradesh, and West Bengal and smaller states and UTs like Sikkim and Goa have low percentages of cooking less houses.

**Insight:**

This implies that the high population areas might experience infrastructural shortcomings in the delivery of simple facilities and therefore policies covering one area at a time to enhance cooking facilities should be implemented especially to states that are in a high deficit.

4.4



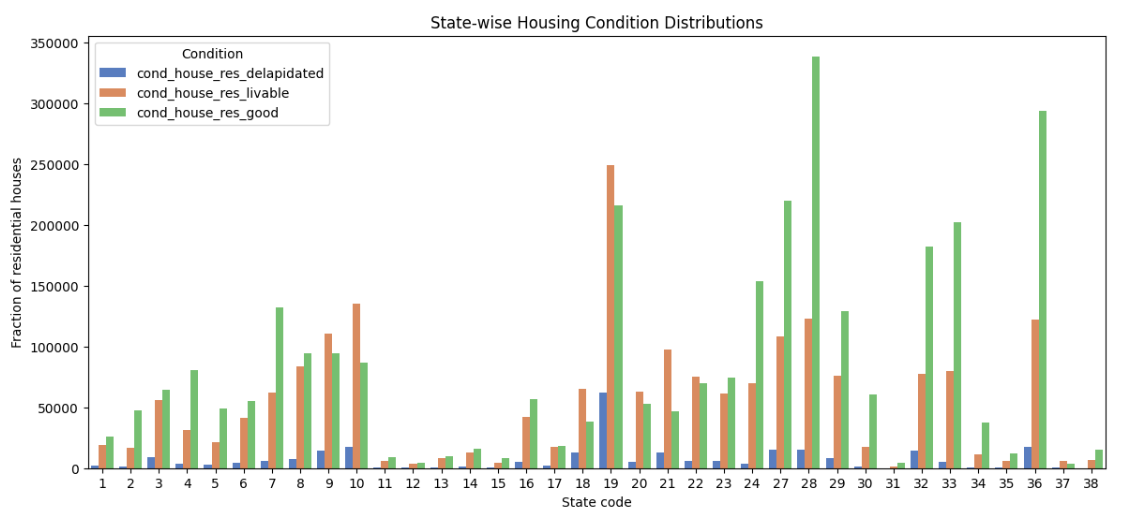
**Observation:**

The scatter plot indicates that the fraction of the number of households using some other fuel is nearly linearly correlated with the fraction of house with no kitchen and this strongly positive relationship is near-perfect.

**Insight:**

This fact means that the use of alternative fuel is also prominent in the regions, where kitchen absence is also highly common pointing to the high correlation between the access to fuel and housing infrastructure.

4.5



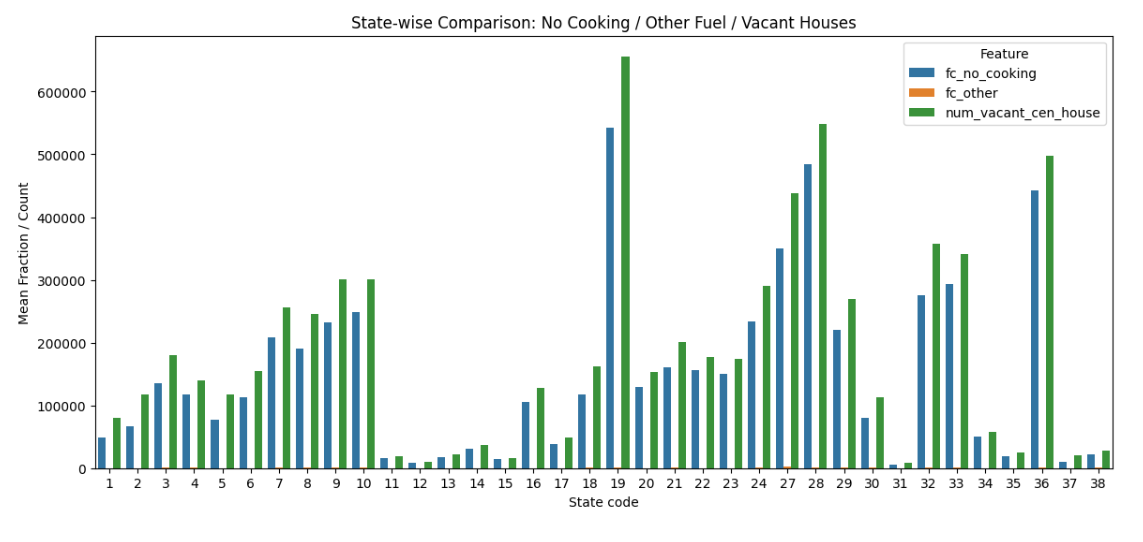
**Observation**

The graph displays the state-wise distribution of residential housing conditions categorized into Good, Livable, and Dilapidated. States like Maharashtra (27), West Bengal (19), and Karnataka (28) have a significantly higher number of good-condition houses compared to other states. However, states such as Uttar Pradesh (9) and Bihar (10) show a notable presence of livable and dilapidated houses, highlighting disparities in housing quality. Smaller states and union territories reflect relatively balanced but lower housing counts across all categories.

**Insight**

This chart underscores a clear housing quality gap between states. While developed states like Maharashtra and Karnataka lead in good-condition housing, states with large populations, such as Uttar Pradesh and Bihar, face quality challenges with a considerable share of substandard housing. This calls for targeted housing improvement programs and infrastructure investments in these regions to enhance living conditions and reduce the reliance on livable or dilapidated housing.

4.6



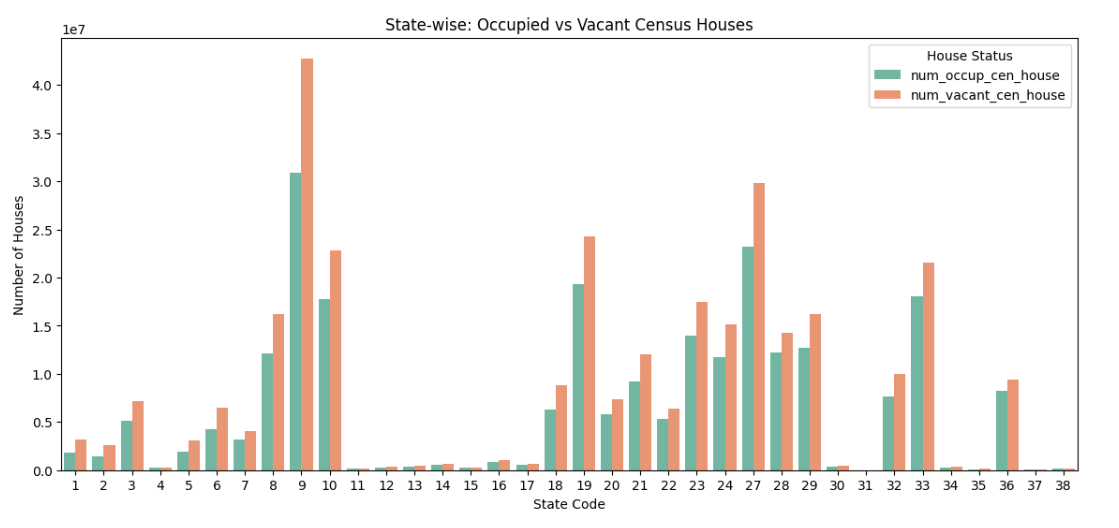
**Observation**

The chart draws a comparison of the three features among states: the houses whose cooking facility is not available (fc\_no\_cooking), the houses that utilized the other types of fuel (fc\_other) and the houses that were unoccupied (num\_vacant\_cen\_house). The high levels of all three attributes as well as deaf houses are reported in such states as West Bengal (19), Maharashtra (27), and Bihar (10). The percentage of the households without proper cooking facilities is also high in states like Uttar Pradesh (9) and Gujarat (24). In the meantime, the values of smaller states and union territories are much lower in all of these categories.

**Insight**

This visualization shows there is a correlation between the housing vacancy and poor amenities which is drastic. Houses which lack proper cooking facility or unsuitable form of fuel in a state whose rate of vacancy is high are also likely to be abandoned thus bad living conditions as a factor of abandonment. Increased access to basic amenities such as kitchen, clean fuel and in states like West Bengal, Maharashtra and Bihar could specifically help to decrease the rate of vacancy and hence more utilisation of housing.

4.7



**Observation**

The chart shows the number of occupied and vacant census of houses in Indian states. The number of houses per unit of housing stock is highest in the Uttar Pradesh State (State Code 9) where the empty houses prevail over occupied ones. Again the same issue of high ratio of housing volume and high level of vacancy is reflected in Maharashtra (27) and Bihar (10). The trend is also great in many other states such as West Bengal (19) and Gujarat (24) that show a high rate of vacancy even though there are a lot of occupied houses. Deeper states and territories of union (codes 12 16 and 35 38) exhibit comparatively low numbers in either case.

**Insight**

The statistics indicate a massive misalignment between housing construction and the households availing of it. States like Uttar Pradesh, Bihar, and Maharashtra have a high level of vacancy which may imply low-quality housing, the cost of housing or inequitable housing supply. In the meanwhile, in smaller states, the occupancy-levels are evened out with the vacancy-levels. To help fill these gaps, there should be policy interventions in the area of affordable housing activities, urban planning investments and revitalization of abandoned houses to make successful use of housing.

4.8



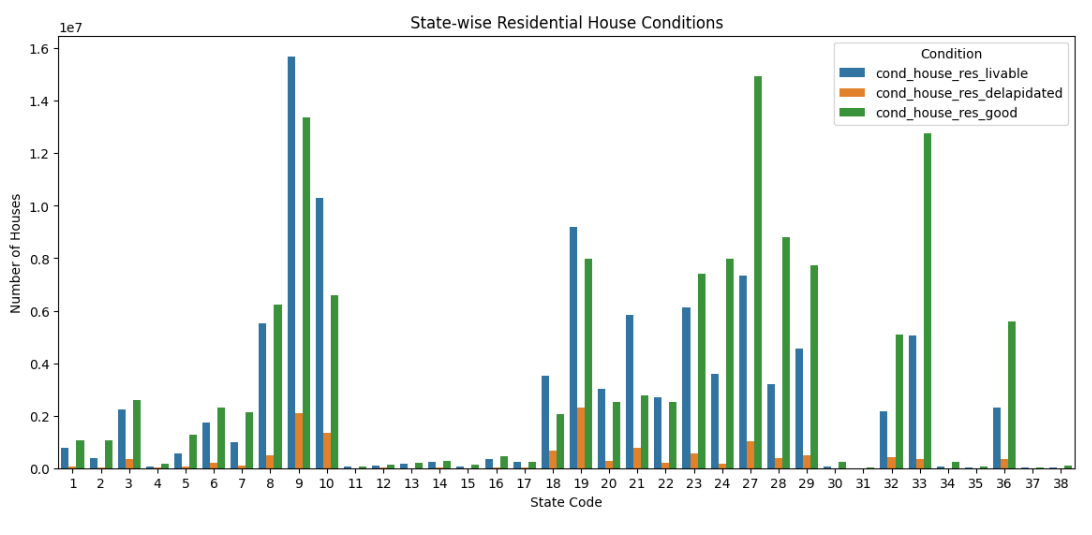
**Observation**

The map presents the allocations of good and poor domestic houses in states of India. The state with the highest number of housing is Uttar Pradesh (State Code 9) where there are a lot of run-down houses and good houses in huge stock. Maharashtra (27) and West Bengal (19) have also good houses in great numbers but this does not mean that there is no significant level of dilapidated housing. Substantially, on the other hand, smaller states and union territories contain fewer houses in both categories based on their size.

**Insight**

Such visualization highlights the discrepancy between the housing quality in high population states. As seen in the case of better-off states like Maharashtra and Gujarat, where the percentage coverage of good-condition houses is higher, the states of Uttar Pradesh and Bihar, to name but a few always seem to trade with a significant amount of deteriorated houses. That is why there is an urgent necessity to carry out renovation and infrastructure improvement programmes in the society where the quality of housing is rather low, even though providing a huge amount of housing. The move on enhancing the quality of housing in these states would enhance a great deal the standards of living and structural risks.

4.9



**Observation**

The bar chart with the inscription State-wise Conditions of residential houses shows the residential houses categorized into three conditions viz. Good, Livable and Dilapidated in terms of their distribution across various Indian States. The most notable finding is that the number of residential houses are higher in State Code 9 (Uttar Pradesh) and a vast platform of residential houses is found out to be categorized as either livable or good. A significant percentage, are however, in the dilapidated category. Close behind in the number of houses, is State Code 10 (Bihar) and State Code 27 (Maharashtra). Maharashtra has a higher density of good houses than that of Bihar. Bihar has a very high percentage of poor quality of houses implying low quality of houses. West Bengal (18), Odisha (21) and Assam (5) also show a large number of living and poor number of dilapidated houses, whereas smaller states and union territories codes 35-38 are expectedly not that many due to the lesser population size

Insight

This is an eye opener kind of a chart with there being a huge difference in the quality of housing infrastructures in the Indian states. States with dense population like U.P., Bihar and West Bengal have a triple burden in that, they have large stocks of houses with a significant proportion of their houses in dilapidated or sub-standard condition, which forebear on the repair and maintenance of their infrastructure. The western states in the country such as Maharashtra and Gujarat exhibit better housing conditions, with larger shares of the housing stock in good conditions. This is illustrative of the greater differences in infrastructure development in the region. The statistics indicate the necessity of specific housing solutions, particularly, in those states where livable houses are predominant and the dilapidated ones. The improvement of current dwellings instead of solely focusing on the quantitative aspect should be included in the agenda of the policy The housing quality must be improved to a great extent to enhance health, safety, and overall well-being of the citizens especially in underserved areas.

### **7. Conclusion**

The Exploratory Data Analysis performed on the Census Household Amenities dataset has provided valuable insights into the housing conditions, basic amenities, and infrastructure disparities across India. The data cleaning and preprocessing steps ensured high data quality, making the analysis robust and reliable for deriving meaningful conclusions.

This EDA has highlighted the following main findings:

* **Infrastructure Gaps in Basic Amenities:** A considerable number of occupied houses lack essential amenities such as proper cooking facilities (fc\_no\_cooking), indicating persistent gaps in housing quality despite overall residential growth.
* **Housing Condition Disparities:** While a majority of houses fall under "livable" or "good" condition categories, the presence of a notable proportion of dilapidated houses (cond\_house\_res\_delapidated) reveals housing distress in several regions, particularly rural districts.
* **Regional Inequalities Across States:** State-wise comparisons show significant differences in access to amenities and housing conditions. Some states exhibit better infrastructure and lower vacancy rates, while others display higher vacancy (num\_vacant\_cen\_house) and weaker housing quality, emphasizing the need for targeted policy interventions.
* **Correlation Between Housing and Socioeconomic Factors:** Strong correlations, such as between fc\_other (use of alternative fuels) and kf\_no\_o\_k (lack of kitchen), indicate that informal or unsafe cooking practices are linked with inadequate housing infrastructure. Similarly, the correlation between total housing stock and vacant houses suggests quantity does not always translate into effective utilization.
* **Data Integrity and Completeness:** The dataset was generally clean with minimal missing data (~0.94%), and systematic handling of these missing values ensured that analysis remained unbiased and representative of national housing conditions.

This EDA serves as a critical analytical resource for understanding India's housing and amenity distribution. These insights can help inform housing policies, improve infrastructure planning, and guide targeted welfare programs aimed at reducing regional disparities and improving living standards across the country.