**Disclaimer:**

Datasets and problem statements, based on real data, has been modified for learning purpose. In no way the data or statements used in the project is to be taken as factual.

**Note:**

* The names of states and union territories may differ in different files due to changes in the spelling or some other reasons. This has to be taken into consideration.
* Some State/UT may be referred to as different spelling in different files, for example
* PONDICHERRY – Puducherry
* ORISSA – Odisha
* NCT OF DELHI – Delhi
* New State/UT created:

Telangana – Separated from the northwestern part of Andhra Pradesh in 2014

* Assume that the data is being analyzed on 1st January 2020.
* Do not change the data in the files in the “Data” folder. Instead, make the required changes in the program and save the final file in the “Clean” folder.

**Clean the Census data**

The healthcare department wants to process the 2011 census data (*Data/census2011.csv*) to find some relevant information about their department.

However, there is a lot of information in the data that is not relevant and can be ignored.

There are differences in the nomenclature in different datasets so a uniform nomenclature needs to be found as well. New states and Union Territories have been formed at the time of analysis which has to be taken care of so that the data can be used with the data that was captured later.

It is also reported that some data is missing in the dataset. However, the values of some of these missing data can be found by using data from other fields.

You have been given the responsibility to address these problems and create clean data that can be used later.

**Problem Statement 1: (Keep the relevant data)**

The census 2011 file contains many fields, which we may not use. Remove some columns so that we are left with only relevant data.

We may need the following columns.

* State name
* District name
* Population
* Male
* Female
* Literate
* Male\_Literate
* Female\_Literate
* Rural\_Households
* Urban\_Households
* Households
* Age\_Group\_0\_29
* Age\_Group\_30\_49
* Age\_Group\_50
* Age not stated

Import the data to pandas and keep only the required columns.

**Problem Statement 2: (Rename the Column names)**

For uniformity in the datasets and taking into consideration the census year, we need to rename some columns.

State name to State/UT

District name to District

Male\_Literate to Literate\_Male

Female\_Literate to Literate\_Female

Rural\_Households to Households\_Rural

Urban\_ Households to Households\_Urban

Age\_Group\_0\_29 to Young\_and\_Adult

Age\_Group\_30\_49 to Middle\_Aged

Age\_Group\_50 to Senior\_Citizen

Age not stated to Age\_Not\_Stated

**Problem Statement 3: (Rename State/UT Names)**

The State/UT names are in all caps in the census data, For uniformity across datasets we use the names so that only the first character of each word in the name is in upper case and the rest are in lower case. However, if the word is “and” then it should be all lowercase.

Examples:

* Andaman and Nicobar Islands
* Arunachal Pradesh
* Bihar

**Problem Statement 4: (New State/UT formation)**

* In 2014 Telangana was formed after it split from Andhra Pradesh, The districts that were included in Telangana are stored in *Data/Telangana.txt* . Read the text file and Rename the State/UT From “Andhra Pradesh” to “Telangana” for the given districts.
* In 2019 Laddakh was formed after it split from Jammu and Kashmir, which included the districts Leh and Kargil. Rename the State/UT From “Jammu and Kashmir” to “Laddakh” for the given districts.

**Problem Statement 5: (Find and process Missing Data)**

Find and store the percentage of data missing for each column.

Some data can be found and filled in by using information from other cells. Try to find the correct data by using information from other cells and filling it in. Find and store the percentage of data missing for each column.

Hint:

* Population = Male + Female
* Literate = Literate\_Male + Literate\_Female
* Population = Young\_and\_Adult+ Middle\_Aged + Senior\_Citizen + Age\_Not\_Stated
* Households = Households\_Rural + Households\_Urban

Create a visualization that compares the amount of missing data before and after the data-filling process was done.

**Problem Statement 6: (Save Data)**

Save the processed data to a CSV file named “*census.csv*” in the *“Clean\_Data*” folder.

**Extract Information from the housing data along with the census data**

Siddhu, from Healthcare Department, has reported that the household data that they have (*Data/housing.csv*), requires some processing in the context of the census data so that more information can be extracted from it.

Process the housing data and census data to gain some insight.

**Problem Statement 7: (Process the relevant data)**

From the housing data(*Data/housing.csv*) read the following columns which may be relevant to the requirement.

* District Name
* Rural/Urban
* Total Number of households
* Total Number of Livable
* Total Number of Dilapidated
* Latrine\_premise

In the given data The total number of households, Total Number of Livable, Total Number of Dilapidated, Latrine\_premise, etc. seem to be in percentage terms and not in absolute terms.

The total number of households in a district can be found in the census data. Take the information from the census data and find the absolute values for the Total Number of Dilapidated, and Latrine\_premise for both rural and urban areas of each district.

Calculate the data and rename the column names and save it in a CSV file named “*housing.csv*” in the *“Clean\_Data*” folder, such that it has the following column names:

* District
* Households\_Rural
* Households\_Rural\_Livable
* Households\_Rural\_Dilapidated
* Households\_Rural\_Toilet\_Premise
* Households\_Urban
* Households\_Urban\_Livable
* Households\_Urban\_Dilapidated
* Households\_Urban\_Toilet\_Premise

Report if any district data is in one file but not the other.

**Problem Statement 8: (Visualise the housing data)**

Visualize the following data for each state using appropriate methods:

* Number of households for 100 people
* Percentage of households that have toilet(s) in premise to the total number of households.
* Urban to rural population ratio.

**Problem Statement 9: (Inconsistency in different datasets)**

After the updates, the **“**Households\_Rural” and “Households\_Urban” data are redundant in both housing and census data. Compare the two data and see if there is a major difference between the values in the two files. A difference of more than 10% can be considered a major difference. Report the names of the districts where a major difference is found in the data.

**Clean and gain Insights from the Hospital Data**

Aliah, as a representative of the department, is willing to know if any insights can be gained from the Hospital data (*Data/hospitals.csv*). Apart from general information she wants to know which states need more hospital beds at the earliest.

Assist Aliah to gain insights from the data after cleaning it.

**Problem Statement 10: (Fix the header)**

Import the hospital data from *Data/hospitals.csv* and make the necessary changes to the header (Do not change the original file).  
The header uses acronyms that are defined in *metadata.csv*. Find the data and rename the headers so that it is more understandable to users who are not familiar with the acronyms. The First cell in the header is missing which should be renamed to State/UT. Rename the other headers in a uniform format.

**Problem Statement 11: (Create a function to alter the data to create uniformity)**

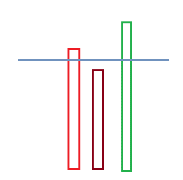
The State/UT names have some different formats compared to other census and housing data. Manually identify those differences and fix them so that the State/UT names match exactly with the census and housing data.

A function should be created to perform this operation. Since the same operation is required for another dataset as well.

After the process save the data in a CSV file named “*all\_hospitals.csv*” in the *“Clean\_Data*” folder

**Problem Statement 12: (Analyze Healthcare facility disparity)**

Visually represent how many hospital beds are there for every 10,000 people in each state or union territory. The national value should also be represented in the same visualization such that the value for each state can be compared to it.

Hint:

Report three States/UTs which have the least amount of beds for their population so that more beds can be added to the hospitals in respective states.

**Government Hospitals Requirement Analysis**

The Union of Doctors in its meeting passed a resolution that the government hospitals are lacking in numbers in many states, The government has taken notice of the resolution and has asked for a recommendation for the state where a new government hospital can be set up.

Use the government hospital data (*Data/government\_hospitals.csv*) to suggest which state the government should create a new hospital.

**Problem Statement 13: (Multi-line header)**

The header in (*Data/government\_hospitals.csv*) is in 2 rows

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| States/UTs | Rural hospitals | | Urban hospitals | | As on |
|  | No. | Beds | No. | Beds |  |

The headers are required to be edited for ease of future use as well as for the sake of uniformity.

Import and update the data in a way that it has the following column names

* State/UT
* Rural\_Government\_Hospitals
* Rural\_Government\_Beds
* Urban\_Government\_Hospitals
* Urban\_Government\_Beds
* Last\_Updated

**Problem Statement 14: (Data update and code reuse)**

The ”Last\_Updated” column contains a date that is in the format DD.MM.YYYY but in the future, the date is required in a different format (YYYY-MM-DD) update the date to the required format.

The State/UT names have some different formats compared to other census and housing data. Fix those issues using the function created earlier. If the function created earlier for all\_hospitals cannot be used for government\_hospital then edit the function so that it can be used for both.

Create a new file *“government\_hospital.csv”* in the *“Clean”* folder and save all the government hospital data after the update.

**Problem Statement 15: (Government healthcare facility disparity)**

Since the resources are limited, it is required to identify the region which lacks the healthcare facility the most for creating new government hospitals.

An idea was suggested\* that first the three States/UTs which have the least amount of beds (in all hospitals government and private) for their population is identified. Among those three the state which has the least number of government hospitals can be recommended for setting up a new government hospital.

**World Health Organization Standards**

According to WHO standards, a minimum of 3 beds per 1000 people is required. Monica wants to analyse whether different regions in India as well as the nation as a whole are up to that standard. If not, how many facilities are required to fill the gap?

**Problem Statement 16: (Gap in number of beds)**

Visually represent the difference between the expected number of hospital beds and the available number of hospital beds in each State/UT, as well as at the national level (if any).

**Problem Statement 17: (Hospitals required to meet the standards)**

In a hypothetical situation if the government wishes to fill the gap between the available facility and the WHO standards all by itself how many government hospitals may be required in each state or union territory as well as at the national level?

Find the average number of beds in a government hospital. Divide it by the gap in the number of beds to reach the standards in the region, and round the number to the nearest integer to find the value.

Represent the findings visually.

**Connecting with Database**

The department of Healthcare wants to access all the clean data to be accessed over the internet through its official website. They have sent a requirement to upload all the data that has been cleaned to a database.

You have been given the responsibility to address their requirement.

**Problem Statement 18: (Database connection and data upload)**

All the files in the clean folder are required to be uploaded to a relational database. The table names should be the same as the file names without the extension.

The primary key and foreign key constraints should be included in the tables wherever required.

**Problem Statement 19: (Verify upload)**

Verification is a necessary process to be followed after uploading the data.

Join all the tables using the respective primary key- foreign key pair and show the top 3 rows.

Verify whether the result is as expected or not. If not, then identify and fix the issue.

**Problem Statement 20: (Run Query on the database and use in python)**

Write a query that finds the number of households without a toilet on the premise for each district of the state that has the least hospital bed-to-population ratio.

Display the information using python.

**Run Queries on a relational database connected to python**

**Problem Statement 21:**

Create stored functions for the following:

|  |  |  |
| --- | --- | --- |
| **Stored Function name** | **Input parameter** | **Return parameter** |
| get\_population\_district | District Name | Total population of the district |
| get\_population | State/UT Name | Total population of the district |
| senior\_citizen\_population | State/UT Name | Total Senior citizens in the state |
| get\_hospital\_beds | State/UT Name | Total Number of hospital beds in the State/UT |
| get\_govt\_hospital\_beds | State/UT Name | Total Number of government hospital beds in the State/UT |
| beds\_per\_lakh | State/UT Name | Total Number of hospital beds per 1 lakh people |
| govt\_beds\_per\_lakh | State/UT Name | Total Number of government hospital beds per 1 lakh people |

**Problem Statement 22:**

It was reported\* that in the north-eastern states, senior citizens are facing some issues in getting beds in a government hospital. Amir, from the department of healthcare, has requested a report on the healthcare situation in the north-eastern states.

The north-eastern states are listed in (*Data/north\_east\_states.txt*). Read the file and extract the state names using python and write a query to find the following information about those states from the database. Use the stored functions created earlier wherever relevant.

* State Name
* Population
* Senior Citizen Population
* Number of Government Hospital
* Number of Government Hospital Beds
* Number of Government Hospital beds for 1 Lakh population

(Round the number to the nearest integer)

* Number of Government Hospital beds for 1 Lakh senior citizen

(Round the number to the nearest integer)

The rows should be sorted by the number of Government Hospital beds for 1 Lakh senior citizens in ascending order.

**Problem Statement 23:**

The Government wants to run a scheme that would help people build in-premise toilets.   
For that purpose, a secretary has requested a district-wise report using the census data.

Using a stored procedure, Create a district-wise report that shows the total population of each district, the total number of households in the district, the number of households that do NOT have on-premise toilets, and the percentage of the number of households that do NOT toilets in premise to the total number of households. Sort the data in a way that is most useful for the secretary.

**Problem Statement 24:**

An Agency wants to find out if there is a relationship between dilapidated homes and the lack of households for the people.

Using a stored procedure, create a report that shows the name of each district, its population, the number of liveable houses per 1000 people, and the number of dilapidated households per 1000 people.

Visualize the relation between these values using an appropriate plot.

**Problem Statement 25:**

The rural development department has requested a report on the healthcare situation in rural areas as compared to urban areas.

Write a query to find the number of hospital beds and government hospital beds per 1 lakh people for rural and urban areas separately for each state, along with the difference between them.

**Problem Statement 26:**

New hospitals are under construction\* in different locations, and once they are constructed the data in the database is to be updated. Also, when some hospitals non-operational hospital might be removed. Since the data is quite significant there should be a hospital\_log table that would store any changes made to the hospital or government\_hospital table.

The hospital\_log table should contain the name of the district, whether the hospital is situated in urban or rural area, whether the hospital is being added or being removed, date of adding or removing the hospital to/from the existing facilities, and whether it is government hospital or not. The a row in the table must be added automatically whenever a new hospital is added, or an existing hospital is removed.

**Problem Statement 27:**

New hospital beds are to be added in different government and private hospitals in different locations, and some non-operational hospital beds might be removed from hospitals. Since the data is quite significant there should be a hospital\_bed\_log table that would store any changes made to the hospital or government\_hospital table with respect to number of hospital beds.

The hospital\_bed\_log table should contain the name of the district, whether the hospital is situated in urban or rural area, date of adding or removing hospital beds, the number of beds being added or removed, whether it is being added or removed and whether it is added to/removed from government hospital or not. The table must be updated automatically whenever new hospital beds are added to or removed from the existing facility.