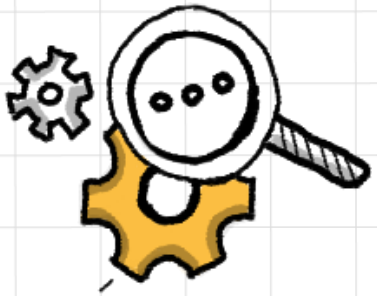
Futurense

Democratizing Tech Talent to deliver impact at scale

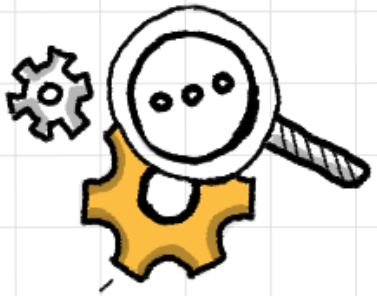
Health Care Project

Dataset link: Datasets



Group F
Team members:

VAIBHAV SIMHA J (GROUP CAPTIAN)
YASH V
REPUDI SAMUEL HONEY
KOTHAPALLI ROHIT CHOWDHARY

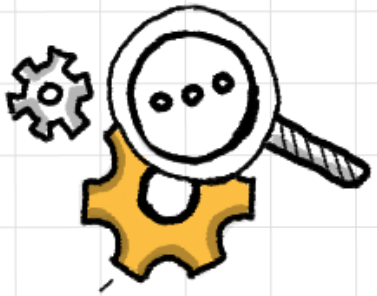


Disclaimer:
Datasets and problem statements, based on real data, has been modified for learning purpose. In no way the data or statements used in the project is to be taken as factual.

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.

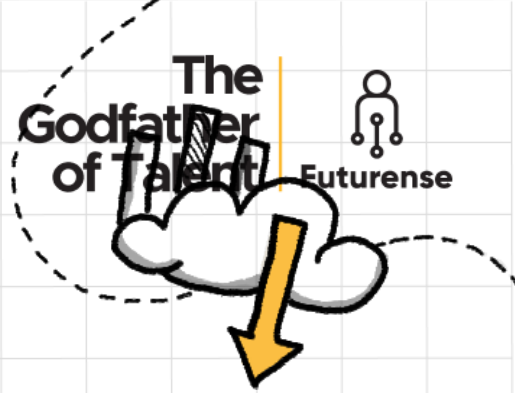
Some State/UT may be referred to as different spelling in different files, for example
PONDICHERRY – Puducherry
ORISSA – Odisha
NCT OF DELHI – Delhi

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

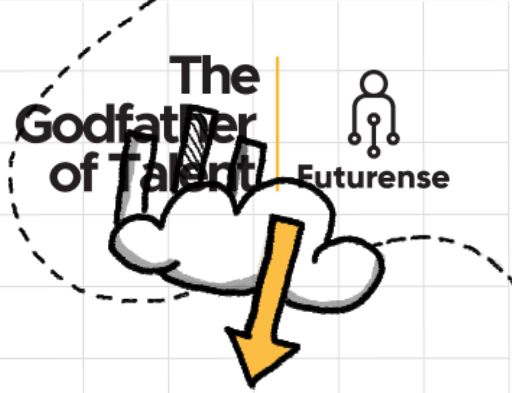
The healthcare department wants to process the 2011 census data (*Data/census2011.csv*) to find some relevant information about their department.

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.

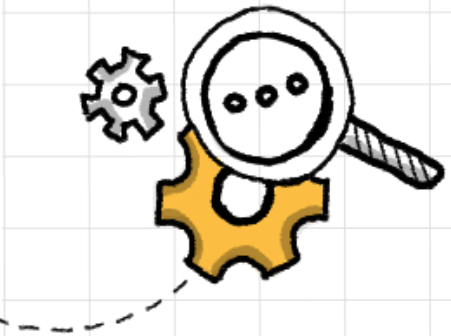


Problem Statement 1: (Keep the relevant data)

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.

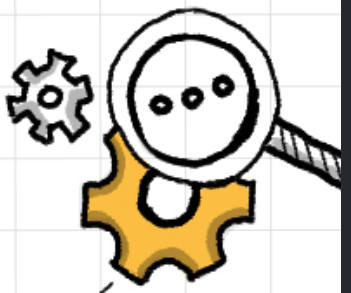
```
# PROBLEM STATEMENT 1 - YASH V
import pandas as pd
census = pd.read_csv('/content/census_2011.csv')
required_columns = ['State name','District name','Population','Male','Female','Literate','Male_Literate','Female_Literate','Rural_Households',
                    'Urban_Households','Households','Age_Group_0_29','Age_Group_30_49','Age_Group_50','Age not stated']
census_selected = census[required_columns]
print(census_selected.head())
```

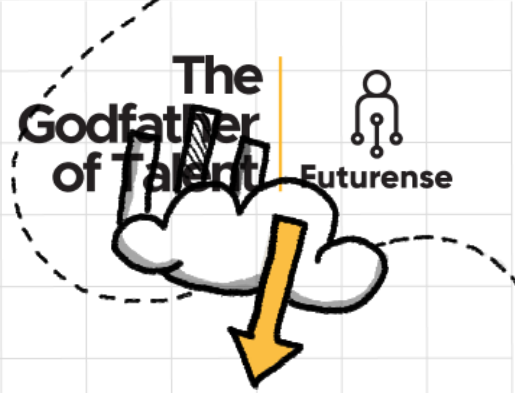
[4]

```
         State name District name  Population      Male    Female  Literate  \
0  JAMMU AND KASHMIR       Kupwara    870354.0  474190.0  396164.0  439654.0
1  JAMMU AND KASHMIR        Badgam    753745.0       NaN  355704.0  335649.0
2  JAMMU AND KASHMIR   Leh(Ladakh)    133487.0   78971.0   54516.0   93770.0
3  JAMMU AND KASHMIR        Kargil    140802.0       NaN   63017.0       NaN
4  JAMMU AND KASHMIR         Punch         NaN  251899.0  224936.0  261724.0

   Male_Literate  Female_Literate  Rural_Households  Urban_Households  \
0       282823.0         156831.0          158438.0               NaN
1       207741.0         127908.0          160649.0           27190.0
2        62834.0          30936.0           36920.0           17474.0
3        56301.0          29935.0           40370.0            7774.0
4       163333.0          98391.0          132139.0           15269.0

   Households  Age_Group_0_29  Age_Group_30_49  Age_Group_50  Age not stated
0    181664.0        600759.0         178435.0       89679.0          1481.0
1    187839.0        503223.0         160933.0       88978.0           611.0
2     54394.0         70703.0          41515.0           NaN           250.0
3     48144.0         87532.0          35561.0       17488.0           221.0
4    147408.0        304979.0         109818.0       61334.0           704.0
```

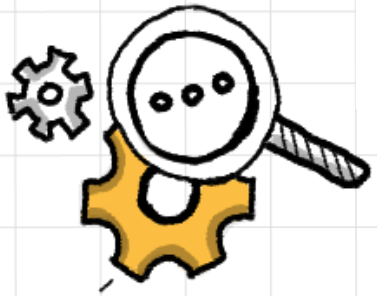


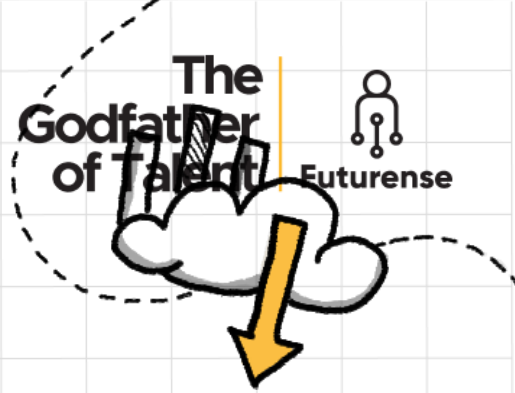


Problem Statement 2: (Rename the Column names)

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



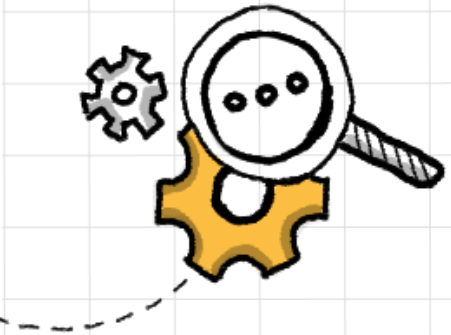


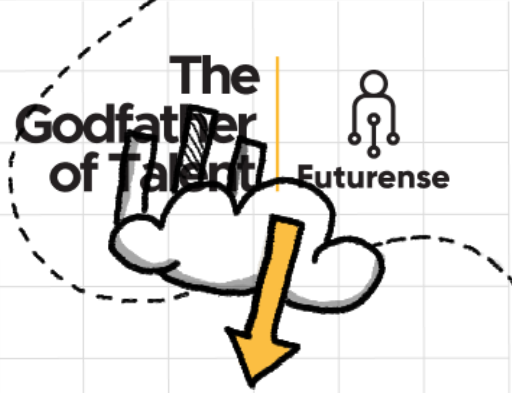

```python
# PROBLEM STATEMENT 2 - VAIBHAV SIMHA J
census_selected = census_selected.rename(columns={
    'State name': 'State/UT',
    'District name': 'District',
    'Male_Literate': 'Literate_Male',
    'Female_Literate': 'Literate_Female',
    'Rural_Households': 'Households_Rural',
    'Urban_Households': 'Households_Urban',
    'Age_Group_0_29': 'Young_and_Adult',
    'Age_Group_30_49': 'Middle_Aged',
    'Age_Group_50': 'Senior_Citizen',
    'Age not stated': 'Age_Not_Stated'
})
census_data=census_selected
print(census_data.head())
```

```
          State/UT     District  Population      Male    Female  Literate  \
0  JAMMU AND KASHMIR      Kupwara    870354.0  474190.0  396164.0  439654.0
1  JAMMU AND KASHMIR       Badgam    753745.0       NaN  355704.0  335649.0
2  JAMMU AND KASHMIR  Leh(Ladakh)    133487.0   78971.0   54516.0   93770.0
3  JAMMU AND KASHMIR       Kargil    140802.0       NaN   63017.0       NaN
4  JAMMU AND KASHMIR        Punch         NaN  251899.0  224936.0  261724.0

   Literate_Male  Literate_Female  Households_Rural  Households_Urban  \
0       282823.0         156831.0          158438.0               NaN
1       207741.0         127908.0          160649.0           27190.0
2        62834.0          30936.0           36920.0           17474.0
3        56301.0          29935.0           40370.0            7774.0
4       163333.0          98391.0          132139.0           15269.0

   Households  Young_and_Adult  Middle_Aged  Senior_Citizen  Age_Not_Stated
0    181664.0         600759.0     178435.0         89679.0          1481.0
1    187839.0         503223.0     160933.0         88978.0           611.0
2     54394.0          70703.0      41515.0             NaN           250.0
3     48144.0          87532.0      35561.0         17488.0           221.0
4    147408.0         304979.0     109818.0         61334.0           704.0
```

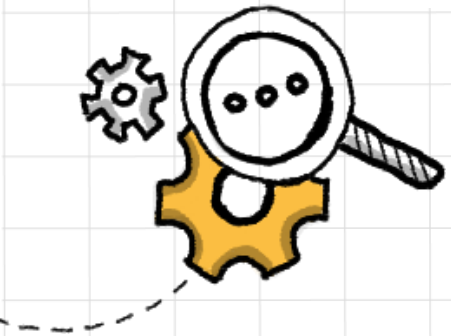


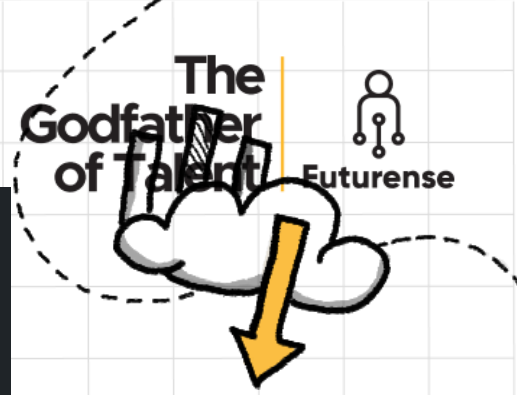


Problem Statement 3: (Rename State/UT Names)

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



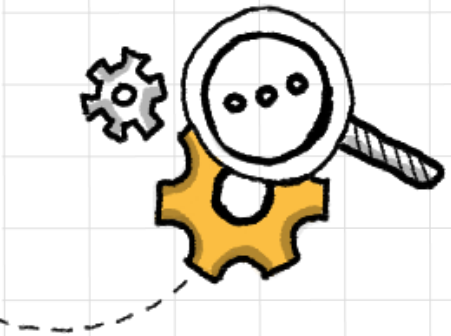


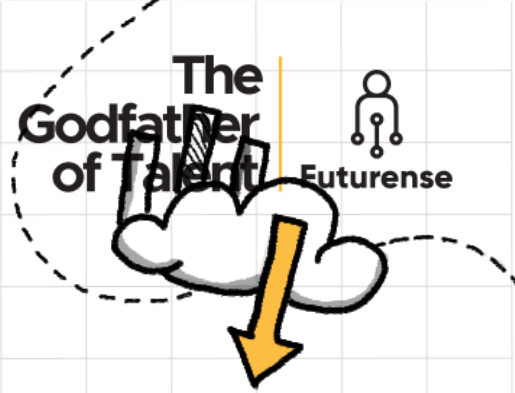

```python
# PROBLEM STATEMENT 3 - REPUDI SAMUEL HONEY
def capitalize_state_name(state):
    words = state.lower().split()
    capitalized_words = [word.capitalize() if word != 'and' else word for word in words]
    return ' '.join(capitalized_words)
census_data['State/UT'] = census_data['State/UT'].apply(capitalize_state_name)
print(census_data.head())
```

```
          State/UT     District  Population      Male    Female  Literate  \
0  Jammu and Kashmir      Kupwara    870354.0  474190.0  396164.0  439654.0
1  Jammu and Kashmir       Badgam    753745.0       NaN  355704.0  335649.0
2  Jammu and Kashmir  Leh(Ladakh)    133487.0   78971.0   54516.0   93770.0
3  Jammu and Kashmir       Kargil    140802.0       NaN   63017.0       NaN
4  Jammu and Kashmir        Punch         NaN  251899.0  224936.0  261724.0

   Literate_Male  Literate_Female  Households_Rural  Households_Urban  \
0       282823.0         156831.0          158438.0               NaN
1       207741.0         127908.0          160649.0           27190.0
2        62834.0          30936.0           36920.0           17474.0
3        56301.0          29935.0           40370.0            7774.0
4       163333.0          98391.0          132139.0           15269.0

   Households  Young_and_Adult  Middle_Aged  Senior_Citizen  Age_Not_Stated
0    181664.0         600759.0     178435.0         89679.0          1481.0
1    187839.0         503223.0     160933.0         88978.0           611.0
2     54394.0          70703.0      41515.0             NaN           250.0
3     48144.0          87532.0      35561.0         17488.0           221.0
4    147408.0         304979.0     109818.0         61334.0           704.0
```

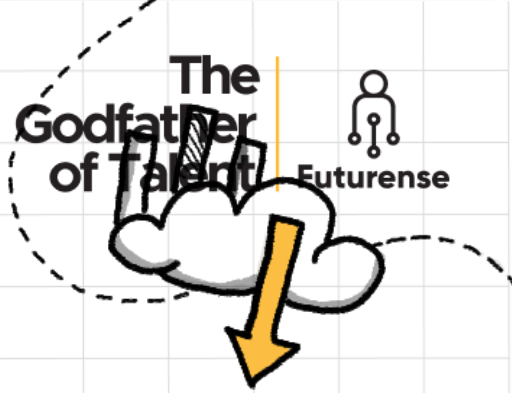




Problem Statement 4: (New State/UT formation)

→ 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 4 - KOTHAPALLI ROHITH CHOWDHARY

    def read_districts_from_file(file_path):
        with open(file_path, 'r') as file:
            return set(line.strip() for line in file)
    df = pd.read_csv('census_2011.csv')

    telangana_districts = set(line.strip() for line in open('/content/Telangana.txt', 'r'))
    df.loc[df['District name'].isin(telangana_districts), 'State name'] = 'Telangana'
    ladakh_districts = {'Leh', 'Kargil'}
    df.loc[df['District name'].isin(ladakh_districts), 'State name'] = 'Ladakh'
    df.to_csv('census_2011_updated.csv', index=False)
    print(df)
```

[9]

```
     District code                  State name              District name  \
0                1           JAMMU AND KASHMIR                     Kupwara
1                2           JAMMU AND KASHMIR                     Badgam
2                3           JAMMU AND KASHMIR                  Leh(Ladakh)
3                4                     Ladakh                     Kargil
4                5           JAMMU AND KASHMIR                      Punch
..             ...                         ...                        ...
635            636                  PONDICHERRY                       Mahe
636            637                  PONDICHERRY                   Karaikal
637            638  ANDAMAN AND NICOBAR ISLANDS                    Nicobars
638            639  ANDAMAN AND NICOBAR ISLANDS  North  AND Middle Andaman
639            640  ANDAMAN AND NICOBAR ISLANDS                South Andaman

     Population      Male    Female  Literate  Male_Literate  Female_Literate  \
0      870354.0  474190.0  396164.0  439654.0       282823.0         156831.0
1      753745.0       NaN  355704.0  335649.0       207741.0         127908.0
2      133487.0   78971.0   54516.0   93770.0        62834.0          30936.0
3      140802.0       NaN   63017.0       NaN        56301.0          29935.0
4           NaN  251899.0  224936.0  261724.0       163333.0          98391.0
```