Section 4517-01 Team 10

Health Analytics among Nine States in India

PARIMALLA POKURI, Northwest Missouri State University, USA
SAI KRISHNA GADDE, Northwest Missouri State University, USA
MAHESH PENDYALA, Northwest Missouri State University, USA
SRILATHA VUNNAM, Northwest Missouri State University, USA
RAJITHA BANDARU, Northwest Missouri State University, USA

Additional Key Words and Phrases: datasets, pyspark, analysis, health

ACM Reference Format:
Parimalla Pokuri, Sai Krishna Gadde, Mahesh Pendyala, Srilatha Vunnam, and Rajitha Bandaru. 2023. Health Analytics among Nine States in India. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 1, Article 1 (April 2023), 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT IDEA

The project idea of the "India - Annual Health Survey(AHS) 2012-13" dataset is to produce a thorough, accurate, and dependable dataset on core vital indicators, including composite ones like Infant Mortality Rate, Maternal Mortality Ratio, and Total Fertility Rate, along with their co-variates (process and outcome indicators), at the district level, and map the changes therein on an annual basis. These benchmarks would aid in a more thorough knowledge and prompt monitoring of the different factors that affect population well-being and health, particularly reproductive and child health. The survey was conducted in the Empowered Action Group (EAG) states Uttarakhand, Rajasthan, Uttar Pradesh, Bihar, Jharkhand,Odisha, Chhattisgarh, Madhya Pradesh and Assam.

2 TECHNOLOGIES USED

- BigData Virtual environment
- Jupyter Notebook
- PySpark
- Python
- SQL Queries
- Overleaf
- Tableau

Authors' addresses: Parimalla Pokuri, Northwest Missouri State University, 8600 Datapoint Drive, Maryville, MO, USA, 64468, S546647@nwmissouri.edu; Sai Krishna Gadde, Northwest Missouri State University, Rono-Hills, Maryville, MO, USA, 64468, S547056@nwmissouri.edu; Mahesh Pendyala, Northwest Missouri State University, 30 Shuangqing Rd, Maryville, MO, USA, 64468, S545506@nwmissouri.edu; Srilatha Vunnam, Northwest Missouri State University, 1 Thørväld Circle, Maryville, MO, USA, 64468, S545534@nwmissouri.edu; Rajitha Bandaru, Northwest Missouri State University, Maryville, MO, USA, 64468, S546904@nwmissouri.edu.

2476-1249/2023/4-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 ARCHITECTURE

- Finalising the dataset
- Importing pandas
- Extracting the csv file for cleaning the unstructured data with pandas.
- Creating PySpark Session, returns a new SparkSession with separate SQLConf, registered temporary views, and UDFs, but shared SparkContext and table cache.
- Reading CSV file using spark
- Writing sql queries using pyspark as per the required goals and it will return output

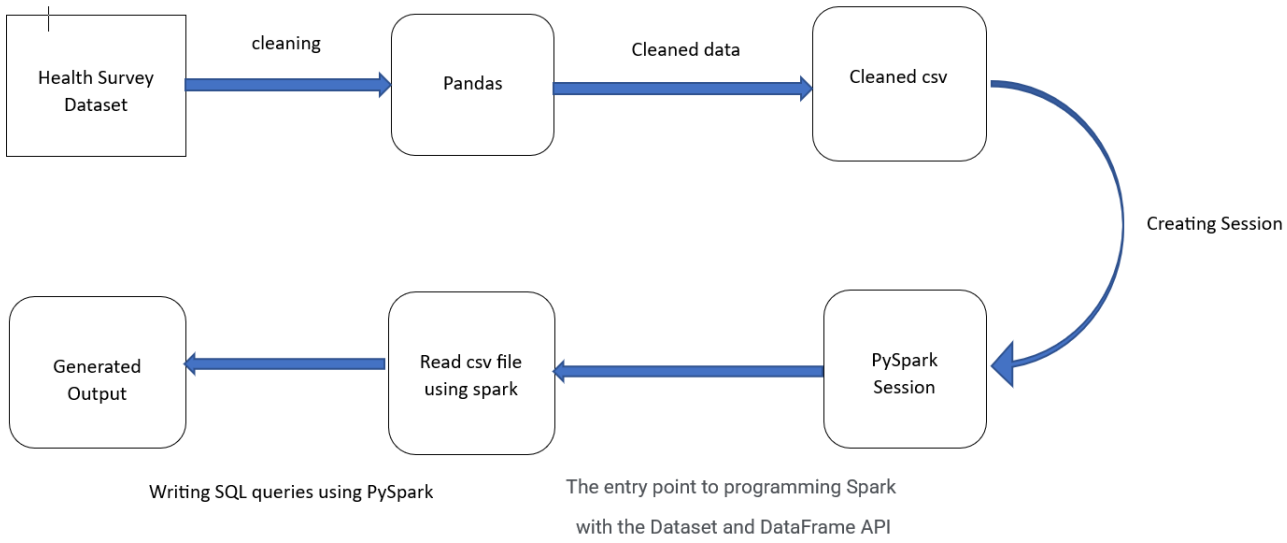

Fig. 1. Data flow of Health Survey

4 GOALS

(1) Percentage of men married before 21 years in Rural and Urban areas
(2) the average population for each state wise
(3) District-wise count of children in rural and urban areas (Ages 12 to 23)
(4) Total number of children currently attending school between 6 to 17 years
(5) Abortion rate in rural and urban areas
(6) Percentage of women who are aware of signs of pneumonia
(7) Which state has the maximum and minimum death rates
(8) Percentage of women married before 18 years in Rural and Urban areas
(9) Which state has max and min birth rates
(10) Usage of copper in both rural and urban areas

5 PROJECT DESCRIPTION

Annual Health Survey (AHS) conducted in India between 2010 and 2013. The AHS is a large-scale survey that collects data on various health indicators from households across India. This dataset includes information on indicators such as maternal and child health, family planning, reproductive health, and utilization of health services. It contains data for over 650 districts across 35 states and union territories in India. The data is provided in the form of CSV files, with each file containing information for a specific year and state/union territory. The dataset includes information on various demographic and socioeconomic factors, such as age, gender, education level, and household income. We can use this dataset to explore trends in health indicators over time and across different regions in India. The dataset can also be used to investigate the impact of various health interventions and policies on health outcomes in different parts of the country.

We can use statistical methods to analyze the data and identify trends and patterns in the health indicators over time and across different regions in India. For example, we could use regression analysis to examine the relationship between various demographic and socioeconomic factors, such as age, gender, education level, and household income, and health outcomes. Health analysis could also involve investigating the impact of various health interventions and policies on health outcomes in different parts of the country. For example, we could compare the health outcomes in states or districts where a particular health intervention has been implemented to those where it has not been implemented to evaluate its effectiveness. We have used PySpark with SQL queries to get previous statistics and predict the future future health outcomes.

Health analysis could involve developing predictive models to forecast future health outcomes based on historical data. These models could help policymakers and healthcare professionals make informed decisions about resource allocation and intervention strategies.

citing here for references [7] [6] [2] [3] [5] [1] [8] [9] [4]

6 RESULTS SUMMARY

- Goal-1: Percentage of men married before 21 years in Rural and Urban areas with respect to the states. We analyzed that in rural areas Rajasthan has a high percentage whereas in urban areas Uttar Pradesh has a high percentage.

Goal-1:

Percentage of men married before 21 years in Rural and Urban areas with respect to state

```
# Maximum Percentage of men married before 21 years in Rural accroding to state
spark.sql('select State_Name,max(EE_Currently_Married_Men_Aged_25_29_Years_Married_Before_Legal_Age_21_Years_Rural)
```

```
+--------------+-------------------------+
|    State_Name|Married_man_before21_rural|
+--------------+-------------------------+
|         Assam|                     32.1|
|         Bihar|                     70.6|
|  Chhattisgarh|                    60.03|
|     Jharkhand|                     49.5|
|Madhya Pradesh|                     68.6|
|        Odisha|                     9.59|
|     Rajasthan|                     78.7|
| Uttar Pradesh|                    64.46|
|   Uttarakhand|                    33.92|
+--------------+-------------------------+
```

Fig. 2. Goal:1

```
]: # Min Percentage of men married before 21 years in Urban area accroding to state
spark.sql('select State_Name,min(EE_Currently_Married_Men_Aged_25_29_Years_Married_Befc
```

```
+--------------+-------------------------+
|    State_Name|Married_man_before21_urban|
+--------------+-------------------------+
|         Assam|                     10.2|
|         Bihar|                     14.7|
|  Chhattisgarh|                    10.48|
|     Jharkhand|                     12.2|
|Madhya Pradesh|                     11.2|
|        Odisha|                    10.81|
|     Rajasthan|                       13|
| Uttar Pradesh|                    14.81|
|   Uttarakhand|                    10.71|
+--------------+-------------------------+
```

Fig. 3. Goal:1

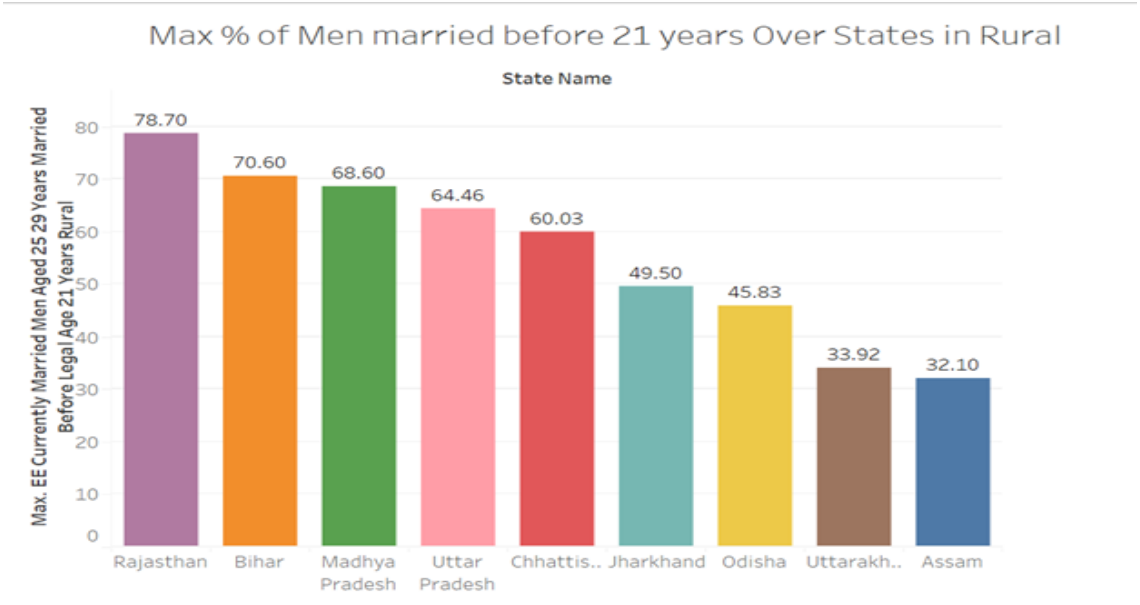

Fig. 4. Visualization for Goal:1

- Goal-2: We have got the total average population for each state wise and we found that Uttarakhand and Madhya Pradesh have the highest and lowest average population respectively.

Goal-2

The total average population for each state wise

```
8]:
spark.sql('select State_Name,avg(AA_Population_Total) as Average_Population from health  group by State_Name o
```

```
+--------------+------------------+
|    State_Name|Average_Population|
+--------------+------------------+
|         Assam| 78678.69565217392|
|         Bihar| 87239.64864864865|
|  Chhattisgarh|        79019.3125|
|     Jharkhand|112183.22222222222|
|Madhya Pradesh| 53106.37777777778|
|        Odisha| 66426.63333333333|
|     Rajasthan|         57128.625|
| Uttar Pradesh|           68692.9|
|   Uttarakhand|132805.92307692306|
+--------------+------------------+
```

Fig. 5. Goal:2

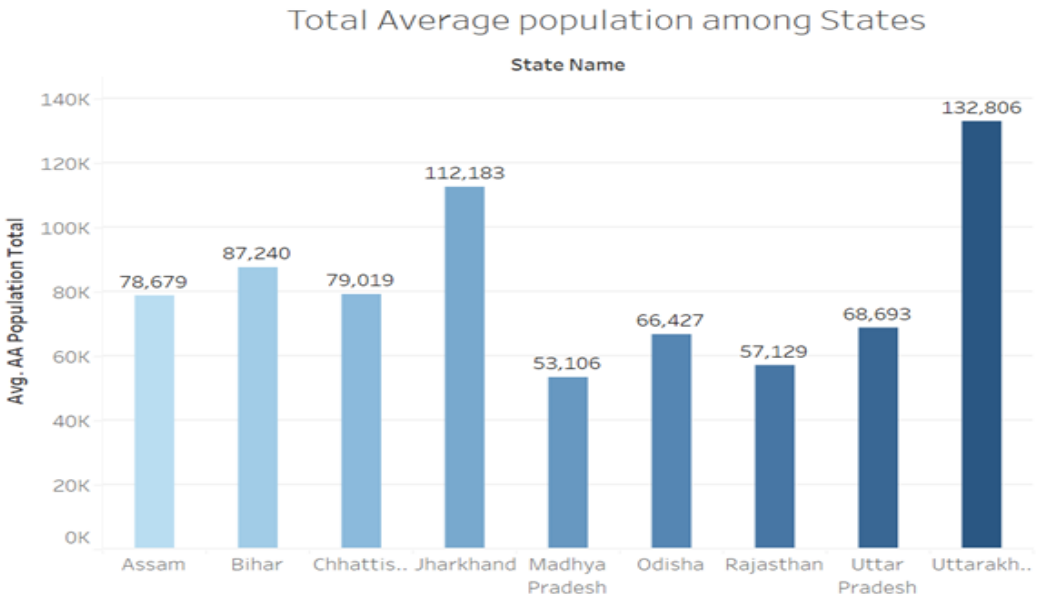

Fig. 6. Visualization for Goal:2

- Goal-3: The number of children aged between 12 to 23 living in rural and urban areas for each district of 7 states in the data set.

Goal3

District-wise count of children in rural and urban areas (Ages 12 to 23)

```
]: spark.sql('select State_Name,State_District_Name, SUM(AA_Children_12_23_Months_Rural) as Rural_Children_Popul
```

```
+--------------+------------------+------------------------+
|    State_Name|State_District_Name|Rural_Children_Population|
+--------------+------------------+------------------------+
|Madhya Pradesh|            Bhopal|                   525.0|
|         Assam|          Kokrajhar|                   566.0|
|         Assam| North Cachar Hills|                   609.0|
|         Assam|         Hailakandi|                   610.0|
|        Odisha|          Rayagada|                   635.0|
|        Odisha|             Baudh|                   681.0|
|Madhya Pradesh|          Jabalpur|                   690.0|
|Madhya Pradesh|           Gwalior|                   699.0|
|        Odisha|          Gajapati|                   722.0|
|        Odisha|         Kandhamal|                   742.0|
|     Rajasthan|           Bikaner|                   749.0|
|         Assam|           Nalbari|                   757.0|
|Madhya Pradesh|             Datia|                   804.0|
```

Fig. 7. Goal:3

```
spark.sql('select State_Name,State_District_Name, SUM(AA_Children_12_23_Months_Urban) as Urban_Children_Population f
```

```
+------------+------------------+------------------------+
|   State_Name|State_District_Name|Urban_Children_Population|
+------------+------------------+------------------------+
|  Uttarakhand|       Rudraprayag|                    20.0|
|        Assam|           Nalbari|                    27.0|
|        Assam|           Darrang|                    34.0|
|       Odisha|      Nabarangapur|                    36.0|
|Uttar Pradesh|        Kushinagar|                    44.0|
|        Assam|         Karimganj|                    47.0|
|        Assam|        Hailakandi|                    50.0|
|        Bihar|         Madhepura|                    52.0|
|       Odisha|           Nuapada|                    52.0|
|Uttar Pradesh|         Shrawasti|                    53.0|
|       Odisha|          Nayagarh|                    54.0|
|    Jharkhand|            Pakaur|                    55.0|
|Uttar Pradesh|         Balrampur|                    57.0|
| Chhattisgarh|           Jashpur|                    57.0|
|        Assam|         Kokrajhar|                    57.0|
|       Odisha|            Baudh|                    58.0|
|        Bihar|        Samastipur|                    68.0|
|        Bihar|         Madhubani|                    68.0|
```

Fig. 8. Goal:3

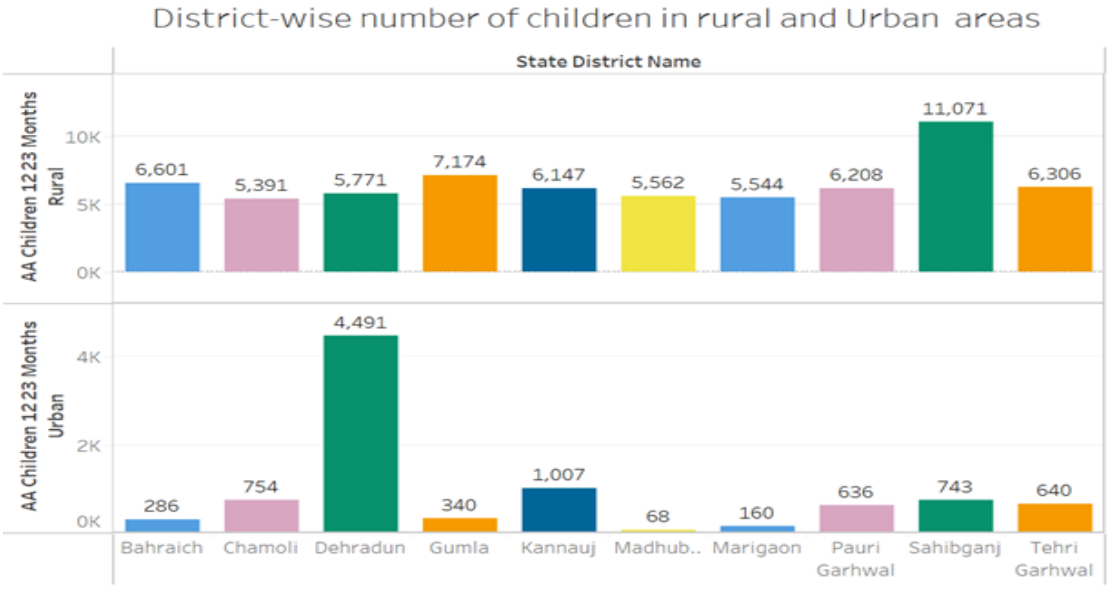

Fig. 9. Visualization for Goal:3

- Goal-4: The count of children aged between 6 to 17 who are enrolled in school and the count of children who did not enroll in school . There can be many reasons, why a child may not be attending school, may include illness, family issues, bullying, learning difficulties, lack of access, and mental health issues

Goal-4

Total number of children currently attending school between 6 to 17 years

```
4]: # FF_Children_Currently_Attending_School_Age_6_17_Years_Person_Total
    # AA_Children_12_23_Months_Total

    spark.sql('select State_District_Name,AA_Children_12_23_Months_Total as Total_Child_Population,round(((FF_Childre
```

State_District_Name	Total_Child_Population	Total_Children_Attending_School	Total_Child_not_attending
Agra	5658	4706.0	952.0
Ajmer	1836	1592.0	244.0
Aligarh	2010	1676.0	334.0
Allahabad	2954	2701.0	253.0
Almora	3664	3601.0	63.0
Alwar	1635	1557.0	78.0
Ambedkar Nagar	1894	1793.0	101.0
Anugul	2448	2051.0	397.0
Araria	3367	3178.0	189.0
Auraiya	3627	3355.0	272.0

Fig. 10. Goal:4

- Goal-5: Average Abortion month by women in rural and urban areas.In general, abortions can be performed during the first trimester (up to 12 weeks), the second trimester (from 13 to 27 weeks), and in rare cases, the third trimester (after 28 weeks). The majority of abortions, around 90 percent, are performed during the first trimester, according to the Centers for Disease Control and Prevention (CDC). Women in rural areas often face greater barriers due to a lack of healthcare providers, transportation, and other factors. Here we found the average month for women who undergo an abortion.

Goal-5

Average Abortion month by women in rural and urban areas

```
#Rural Areas

spark.sql('select State_Name,round(avg(MM_Average_Month_Of_Pregnancy_At_The_Time_Of_Abortion_Rural),1) as avg_aborti
```

```
+--------------+------------------------+
|    State_Name|avg_abortion_month_rural|
+--------------+------------------------+
|         Assam|                     2.7|
|         Bihar|                     3.1|
|  Chhattisgarh|                     3.5|
|     Jharkhand|                     3.0|
|Madhya Pradesh|                     3.4|
|        Odisha|                     3.3|
|     Rajasthan|                     3.1|
| Uttar Pradesh|                     2.9|
|   Uttarakhand|                     3.1|
+--------------+------------------------+
```

Fig. 11. Goal:5

```
#Urban Areas

spark.sql('select State_Name,round(avg(MM_Average_Month_Of_Pregnancy_At_The_Time_Of_Abortion_Urban),1) as avg_aborti
```

```
+--------------+------------------------+
|    State_Name|avg_abortion_month_urban|
+--------------+------------------------+
|         Assam|                     2.6|
|         Bihar|                     2.8|
|  Chhattisgarh|                     3.0|
|     Jharkhand|                     2.7|
|Madhya Pradesh|                     3.2|
|        Odisha|                     3.0|
|     Rajasthan|                     2.8|
| Uttar Pradesh|                     2.7|
|   Uttarakhand|                     2.7|
+--------------+------------------------+
```

Fig. 12. Goal:5

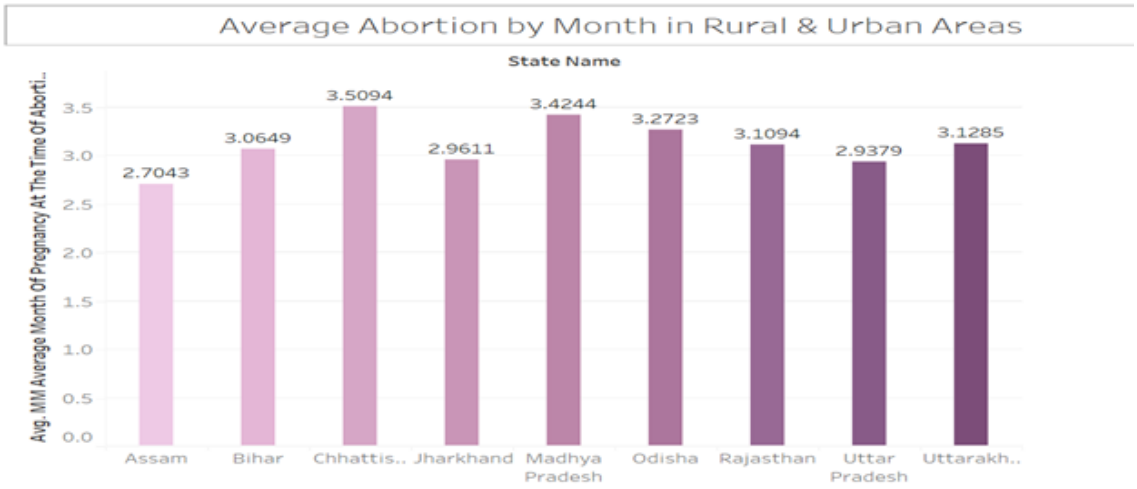

Fig. 13. Visualization for Goal:5

- Goal-6: Percentage of women who are aware of signs of pneumonia, Pneumonia is an infection that affects one or both lungs. Pneumonia is a serious infection that can be caused by bacteria, viruses, or other microorganisms, and it can affect people of all ages, although it's more common among young children, older adults, and people with weakened immune systems. The symptoms of pneumonia can vary, but they often include cough, fever, shortness of breath, chest pain, and fatigue.
  Healthcare providers can play an important role in educating their patients about the signs of pneumonia and encouraging them to seek care if they develop symptoms. Here we have analysed that Odisha has less number of people who are aware of pneumonia.

Goal-6

Percentage of women who are aware of signs of pneumonia

```
]: # XX_Women_Who_Are_Aware_Of_Danger_Signs_Of_Ari_Pneumonia_Total
spark.sql('select State_Name,round(avg(XX_Women_Who_Are_Aware_Of_Danger_Signs_Of_Ari_Pneumonia_Total),2) as avg_per
```

```
+--------------+----------------------------------------+
|    State_Name|avg_percentage_women_aware_about_pneumonia|
+--------------+----------------------------------------+
|         Assam|                                   85.27|
|         Bihar|                                   96.62|
|  Chhattisgarh|                                   91.84|
|     Jharkhand|                                   87.37|
|Madhya Pradesh|                                   94.46|
|        Odisha|                                   59.67|
|     Rajasthan|                                   94.39|
| Uttar Pradesh|                                   95.75|
|   Uttarakhand|                                   96.63|
+--------------+----------------------------------------+
```

Fig. 14. Goal:6

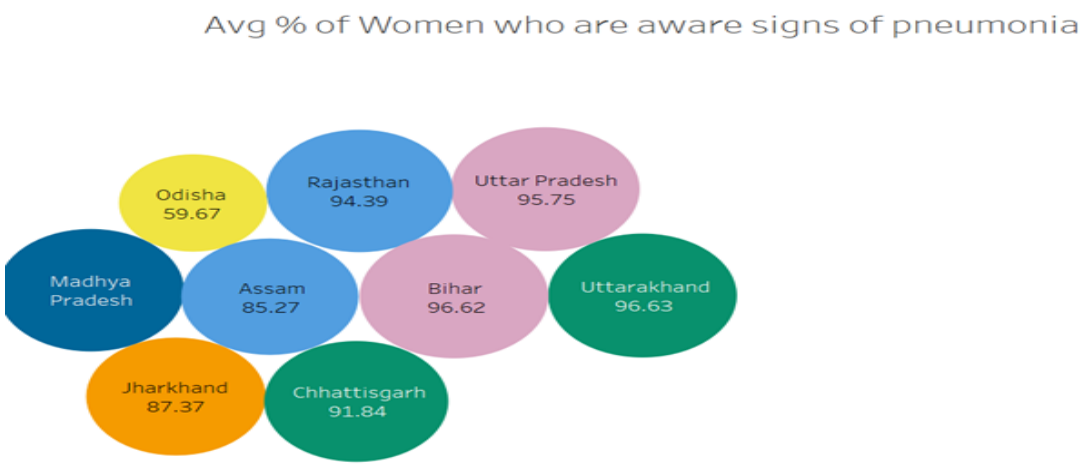

Fig. 15. Visualization for Goal:6

- Goal-7: Which state has the maximum and minimum death rates. The death rate can be affected by a variety of factors, such as age, sex, race, ethnicity, and underlying health conditions. In general, death rates tend to be higher among older adults, males, and people with chronic health conditions, although there can be significant variations within and among different populations. Government agencies and other organizations often collect and publish data on death rates, which can be used to inform public health policies and interventions to improve health outcomes. These sources can provide more detailed information on death rates by cause of death, age group, geographic region, and other factors.

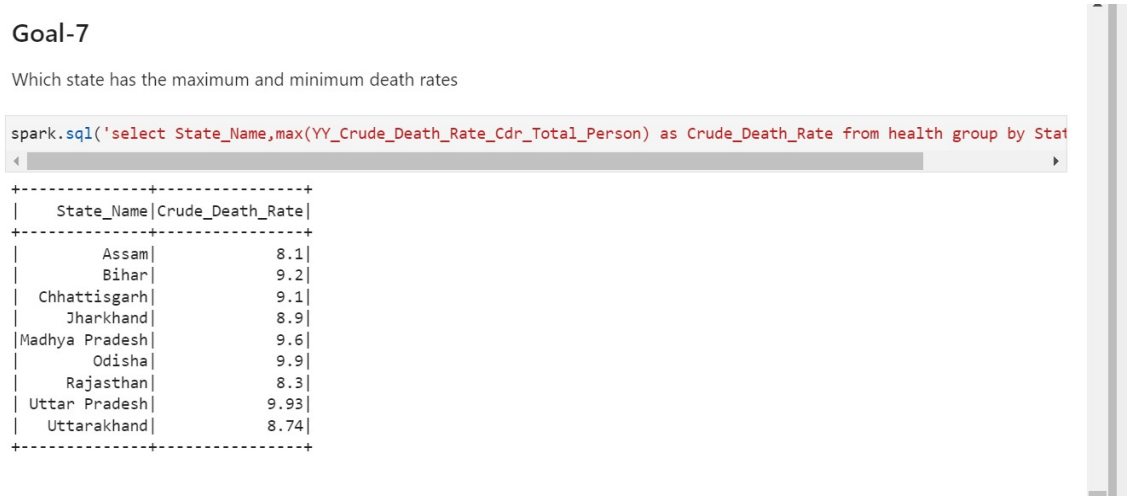

Fig. 16. Goal:7

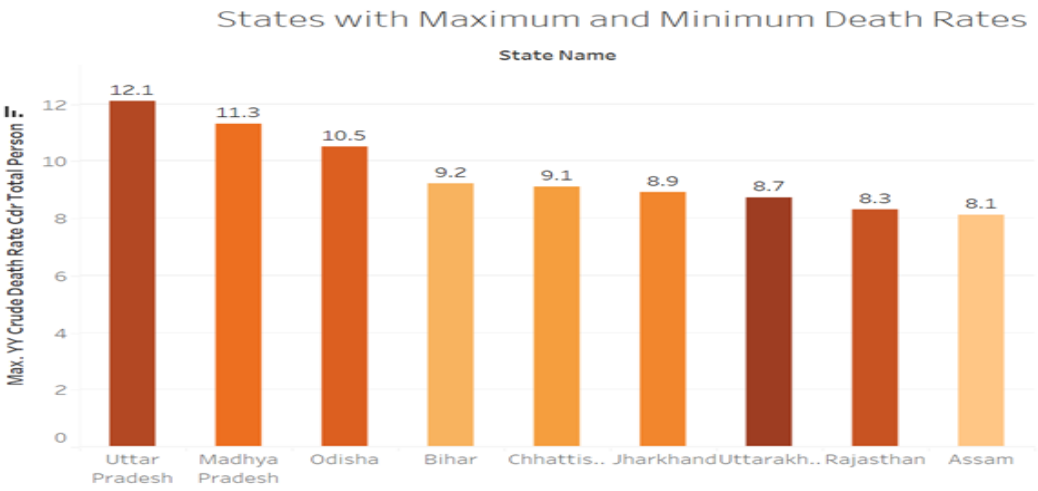

Fig. 17. Visualization for Goal:7

- Goal-8: Percentage of women married before 18 years in Rural and Urban areas. Child marriage refers to a formal or informal union in which one or both parties are under the age of 18. It is a violation of human rights and can have serious consequences for the health and well-being of girls, including increased risk of maternal mortality, domestic violence, and limited educational and economic opportunities. The prevalence of child marriage can vary widely by country, region, and community, and can be influenced by factors such as poverty, gender inequality, and cultural norms. In general, child marriage rates tend to be higher in rural areas and among certain ethnic or religious groups, although there can be significant variations within and among different populations.

Goal-8

Percentage of women married before 18 years in Rural and Urban areas

```
spark.sql('select State_Name,round(avg(EE_Marriages_Among_Females_Below_Legal_Age_18_Years_Total),2) as women_ma
```

```
+-------------+-----------------+
|   State_Name|women_married_b4_18|
+-------------+-----------------+
|  Uttarakhand|             2.13|
| Chhattisgarh|             4.57|
|       Odisha|             4.93|
|Uttar Pradesh|             6.15|
|        Assam|              7.5|
|Madhya Pradesh|            10.98|
|    Jharkhand|            11.98|
|        Bihar|             14.7|
|    Rajasthan|            15.03|
+-------------+-----------------+
```

Fig. 18. Goal:8

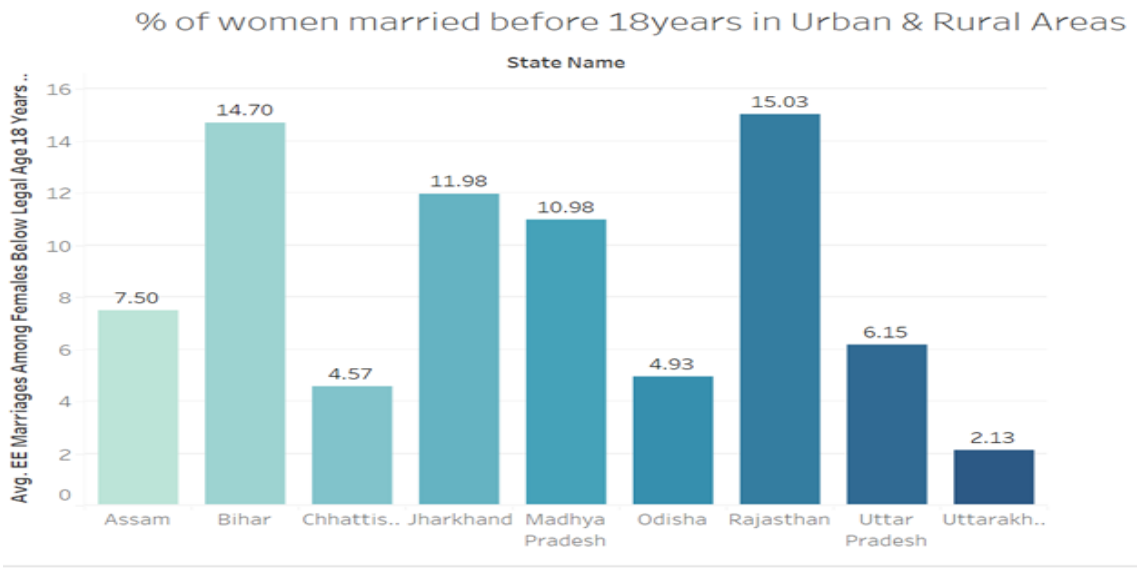

Fig. 19. Visualization for Goal:8

- Goal-9: Which state has max and min birth rates, Birth rates are a key factor in determining population growth. High birth rates can lead to rapid population growth, which can strain resources and put pressure on infrastructure, healthcare systems, and other public services. Low birth rates can result in population decline, which can have implications for workforce participation, economic growth, and social programs.

Goal-9

Which state has max and min birth rates

```
# spark.sql('select round(max(LL_Crude_Birth_Rate_Cbr_Total),2) as Max_birth_rate from health order by 1').show()

# SELECT State_Name, LL_Crude_Birth_Rate_Cbr_Total
# FROM health
# WHERE (LL_Crude_Birth_Rate_Cbr_Total) IN (
#     SELECT MAX(LL_Crude_Birth_Rate_Cbr_Total)
#     FROM health
#     GROUP BY State_Name
# )
spark.sql('SELECT State_Name, LL_Crude_Birth_Rate_Cbr_Total FROM health WHERE (State_Name,LL_Crude_Birth_Rate_Cbr_Total) IN (SELECT Sta
```

```
+--------------+----------------------------+
|    State_Name|LL_Crude_Birth_Rate_Cbr_Total|
+--------------+----------------------------+
|         Assam|                        30.6|
|         Bihar|                        31.2|
|  Chhattisgarh|                       29.43|
|     Jharkhand|                        29.2|
|Madhya Pradesh|                        31.3|
|        Odisha|                        29.1|
|     Rajasthan|                        31.8|
| Uttar Pradesh|                       39.89|
|   Uttarakhand|                       21.94|
```

Fig. 20. Goal:9

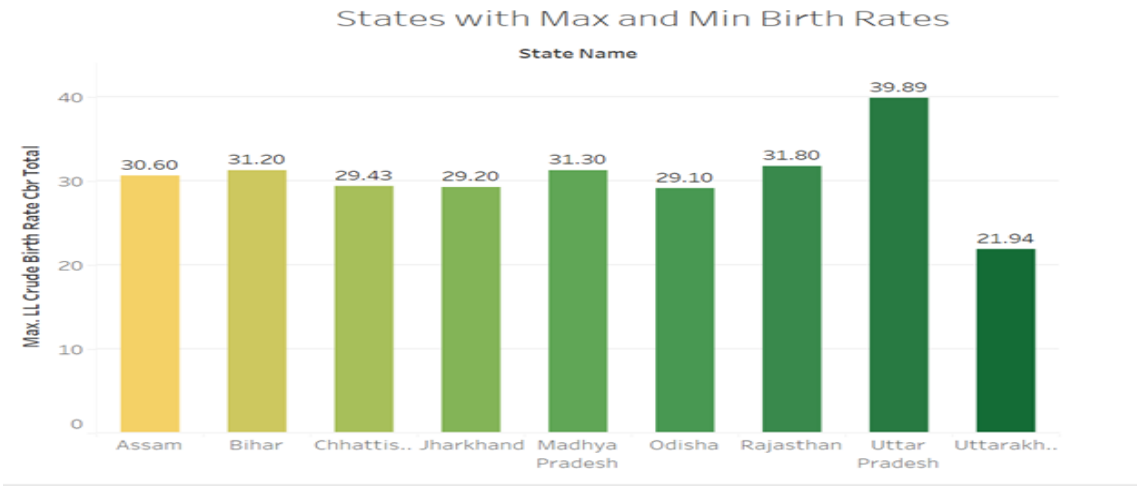

Fig. 21. visualization of Goal:9

- Goal-10: Usage of copper in both rural and urban areas. In both rural and urban areas, copper cookware is commonly used for its heat conductivity and even heat distribution. However, copper cookware may also offer health benefits, as copper ions released during cooking may have antimicrobial properties that can help to prevent foodborne illnesses. While copper has been used for centuries for its health benefits, it is important to note that excessive consumption of copper can be toxic and may lead to health problems. It is also important to ensure that copper is sourced and used in a safe and sustainable manner to avoid environmental damage and human health risks.

Goal-10

Usage of copper in both rural and urban areas

```
# NN_Current_Usage_Copper_T_Iud_Rural
# NN_Current_Usage_Copper_T_Iud_Urban

spark.sql('select State_Name,round(avg(NN_Current_Usage_Copper_T_Iud_Rural),2) as copper_usage_rural from health group by State_Na
```

```
+--------------+------------------+
|    State_Name|copper_usage_rural|
+--------------+------------------+
|Madhya Pradesh|              0.24|
|        Odisha|               0.3|
|     Jharkhand|              0.55|
|  Chhattisgarh|              0.55|
|         Bihar|              0.61|
|   Uttarakhand|              0.62|
|     Rajasthan|              0.98|
| Uttar Pradesh|              1.03|
|         Assam|              1.79|
+--------------+------------------+
```

Fig. 22. Goal:10

```
spark.sql('select State_Name,round(avg(NN_Current_Usage_Copper_T_Iud_urban),2) as copper_usage_urban from health group by Stat
```

```
+--------------+------------------+
|    State_Name|copper_usage_urban|
+--------------+------------------+
|         Assam|              1.14|
|         Bihar|              1.24|
|  Chhattisgarh|              0.93|
|     Jharkhand|              1.31|
|Madhya Pradesh|              0.52|
|        Odisha|              0.74|
|     Rajasthan|              1.96|
| Uttar Pradesh|              1.79|
|   Uttarakhand|               1.5|
+--------------+------------------+
```

Fig. 23. Goal:10

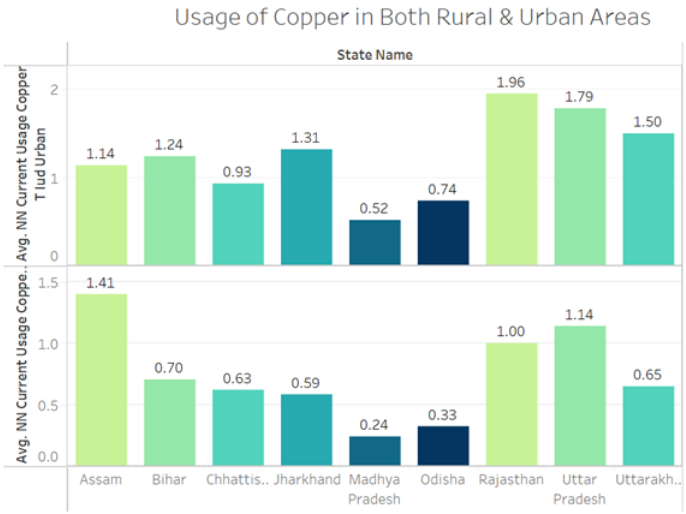

Fig. 24. visualization of Goal:10

7 CONCLUSION

To Summarize, Annual Health Survey is to produce a thorough, accurate, and reliable dataset on the most important health indicators, such as composite ones like infant mortality rate, maternal mortality ratio, and total fertility rate, as well as their covariates (process and outcome indicators) at the district level, and to map annual changes in those indicators. These benchmarks will aid in a more thorough understanding of the many factors that affect population health and well-being, particularly reproductive and child health, as well as timely monitoring of those factors.

REFERENCES

[1] A. M. Alayba, V. Palade, M. England, and R. Iqbal. Arabic language sentiment analysis on health services. In *2017 1st international workshop on arabic script analysis and recognition (asar)*, pages 114–118. IEEE, 2017.

[2] C. Ardington and B. Gasealahwe. Health: analysis of the nids wave 1 and 2 datasets. 2012.

[3] M. J. Bolland, A. Grey, A. Avenell, G. D. Gamble, and I. R. Reid. Calcium supplements with or without vitamin d and risk of cardiovascular events: reanalysis of the women's health initiative limited access dataset and meta-analysis. *Bmj*, 342, 2011.

[4] S. L. Botman and S. S. Jack. Combining national health interview survey datasets: issues and approaches. *Statistics in medicine*, 14(5-7):669–677, 1995.

[5] T. Lumley, P. Diehr, S. Emerson, and L. Chen. The importance of the normality assumption in large public health data sets. *Annual review of public health*, 23(1):151–169, 2002.

[6] C. J. Murray, A. D. Lopez, R. Black, R. Ahuja, S. M. Ali, A. Baqui, L. Dandona, E. Dantzer, V. Das, U. Dhingra, et al. Population health metrics research consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets. *Population health metrics*, 9(1):1–15, 2011.

[7] R. D. Riley, J. Ensor, K. I. Snell, T. P. Debray, D. G. Altman, K. G. Moons, and G. S. Collins. External validation of clinical prediction models using big datasets from e-health records or ipd meta-analysis: opportunities and challenges. *bmj*, 353, 2016.

[8] J. Robertson, P. McElduff, S.-A. Pearson, D. A. Henry, K. J. Inder, and J. R. Attia. The health services burden of heart failure: an analysis using linked population health data-sets. *BMC health services research*, 12(1):1–11, 2012.

[9] M. E. Seligman, C. Peterson, A. J. Barsky, J. K. Boehm, L. D. Kubzansky, N. Park, and D. Labarthe. Positive health and health assets: Re-analysis of longitudinal datasets. 2013.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009