3 GOOD HEALTH AND WELL-BEING

Understanding secondary factors contributing to child mortality

CSE 545 - Big Data Analytics (Spring `20)

Team XYZ - Karan Shah, Komal Agarwal, Narayan Acharya, Vikas Dhyani

3 GOOD HEALTH AND WELL-BEING

What are we doing?

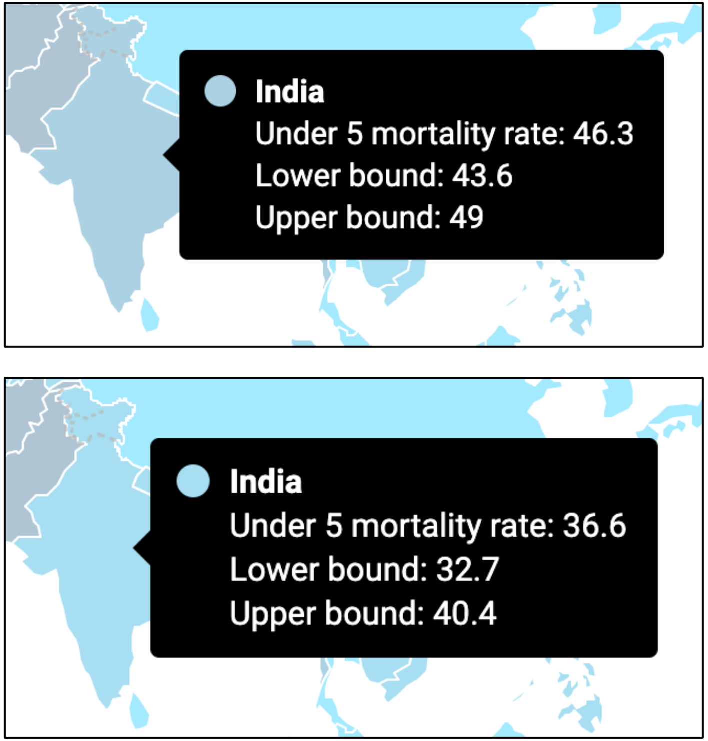
We look to **study the impact of secondary factors** that can help decrease child mortality rate: Environment, Amenities, Lifestyle, Health. And, we try to **suggest priority actions** for Government bodies to action.

Why should one care?

- "...**52 million children under 5 years of age, will die between 2019 and 2030**." *(Source: UNICEF)*
- Even though we are seeing progress it is NOT enough to control the disturbing projections we see for the decade.

How does it relate to SDG?

- SDG 3 -- **"Ensure healthy lives and promoting well-being for all at all ages."**
- Containing infant mortality helps mental and physical well-being of mother and child helping them live healthy lives.

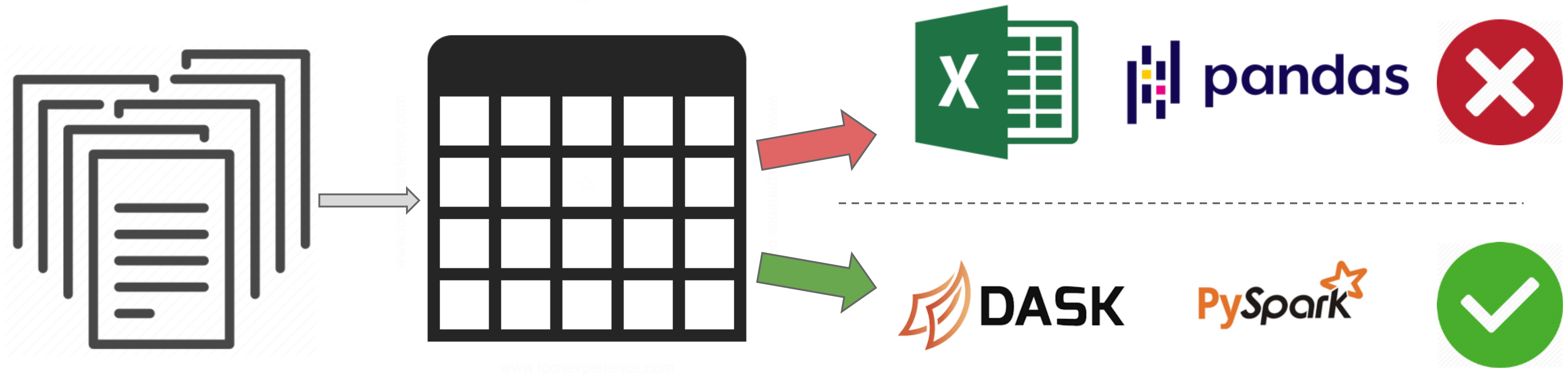


● **India**
Under 5 mortality rate: 46.3
Lower bound: 43.6
Upper bound: 49

● **India**
Under 5 mortality rate: 36.6
Lower bound: 32.7
Upper bound: 40.4

Under 5 mortality rates for India
in 2014 (top) vs 2018 (bottom) from UNICEF data.

Why Big Data?



RAW DATASET:
16.81 GB

> 3M Records
> 200 Columns/Record

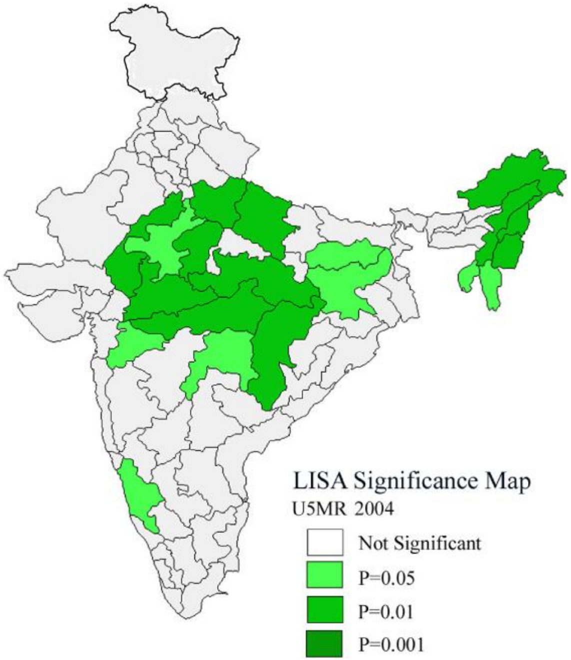
Traditional Programs suffer when entire data is not in memory.

Traditional Programs do NOT support distributed and parallel processing out of the box.

3 GOOD HEALTH AND WELL-BEING

Background

- A Geospatial Analysis study (2011) has shown intra-state and inter-regional disparities in infant mortality in India with help of geospatial techniques like,
  - Moran's-*I*
  - univariate LISA
  - bivariate LISA
  - Spatial error regression
  - spatiotemporal regression



LISA (Cluster and Significance) map depicting spatial clustering and spatial outliers of under-five mortality by incidence of poverty across 74 geographic regions

3 GOOD HEALTH AND WELL-BEING

Background

What is exactly a Infant Mortality Rate?

Number of deaths of infants (age under one year) in 1,000 live births in a year.

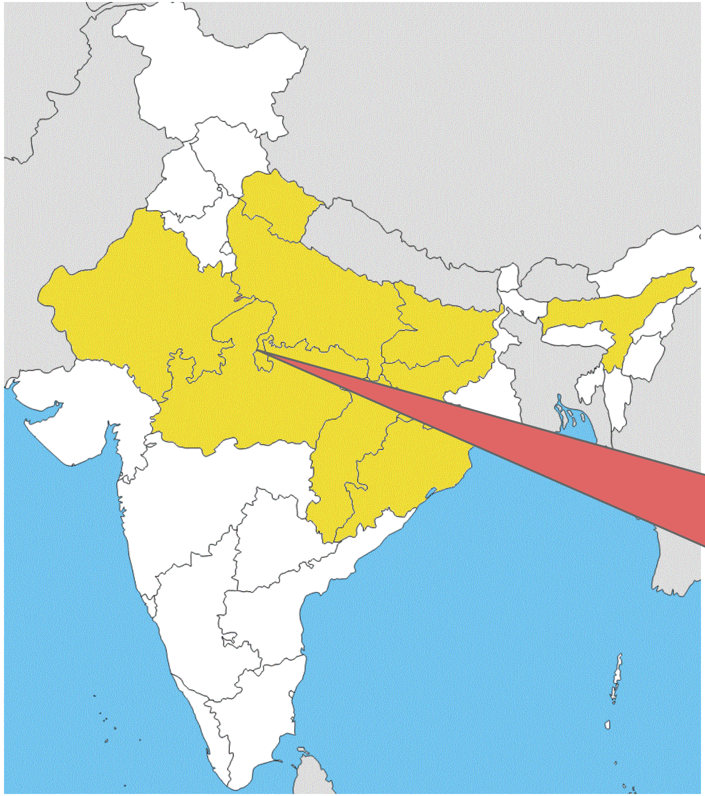
Live births? How is it different than just.... "births"?

Stillbirth - death or loss of a baby before or during delivery

Live birth - baby's birth, showing any sign of life after exiting maternal body

3 GOOD HEALTH AND WELL-BEING

Data



children_lost

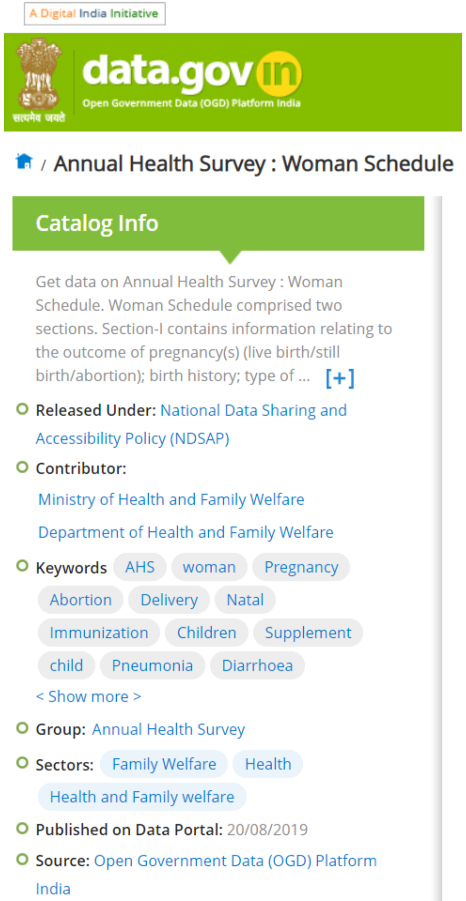
(label)

=

born_alive_total - surviving_total

(features)

- 9 states of India
- ~16.81 GB size
- 200+ features (environment, amenities, lifestyle, health,..)
- 3M records (each record for a woman, district wise)



A Digital India Initiative

data.gov.in

Open Government Data (OGD) Platform India

सत्यमेव जयते

/ Annual Health Survey : Woman Schedule

Catalog Info

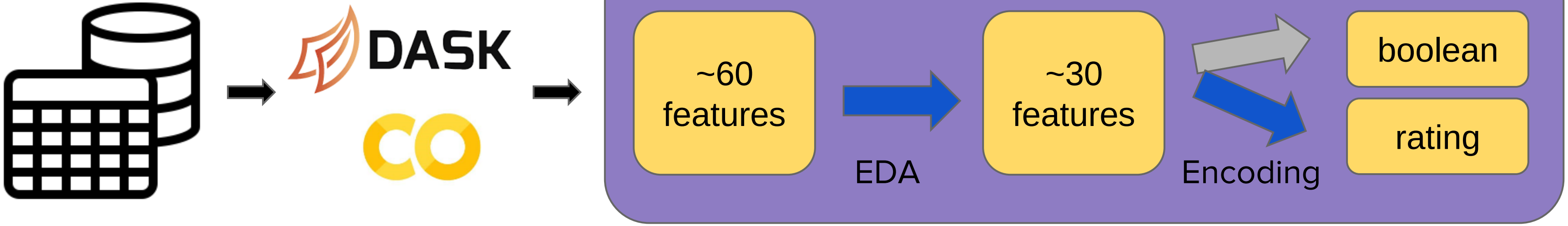
Get data on Annual Health Survey : Woman Schedule. Woman Schedule comprised two sections. Section-I contains information relating to the outcome of pregnancy(s) (live birth/still birth/abortion); birth history; type of ... [+]

- **Released Under:** National Data Sharing and Accessibility Policy (NDSAP)
- **Contributor:**
  Ministry of Health and Family Welfare
  Department of Health and Family Welfare
- **Keywords** AHS woman Pregnancy Abortion Delivery Natal Immunization Children Supplement child Pneumonia Diarrhoea
  < Show more >
- **Group:** Annual Health Survey
- **Sectors:** Family Welfare Health Health and Family welfare
- **Published on Data Portal:** 20/08/2019
- **Source:** Open Government Data (OGD) Platform India

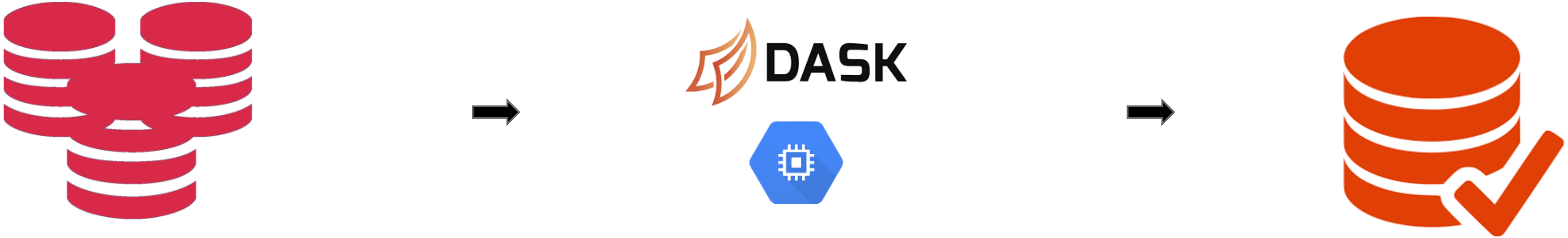
Method

1) Preprocessing, EDA

○ Based on EDA on 1 state data using Google Colaboratory,



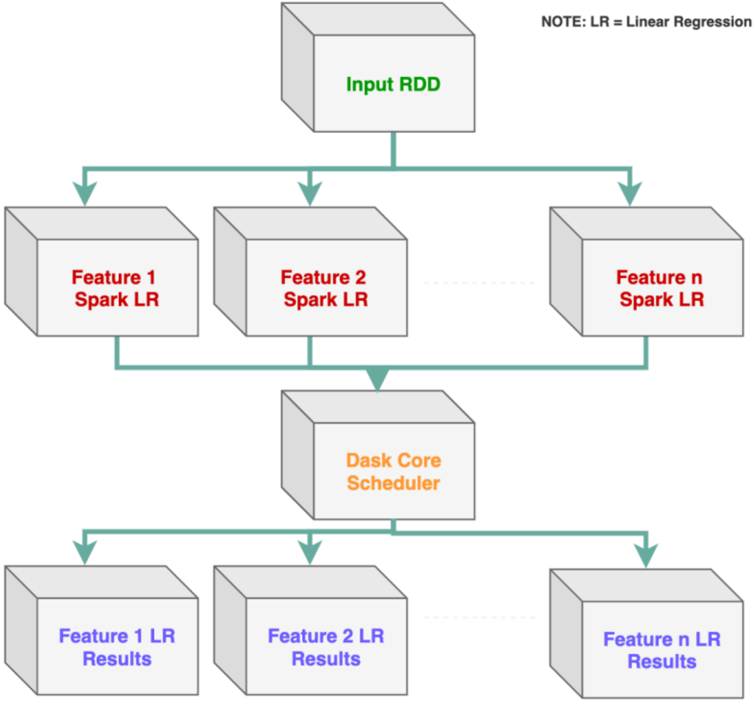
○ Pipeline created to generate full data on GCP cluster from all 9 states raw data



Linear Regression

- We had close to 30 features for independent linear regressions.
  - Dask for parallel processing of multiple independent univariate linear regressions.
  - Individual linear regressions were computed using Spark.

- We used only Spark for Multivariate Linear Regression.
  - Since there was only one linear regression computation, Dask was not needed.

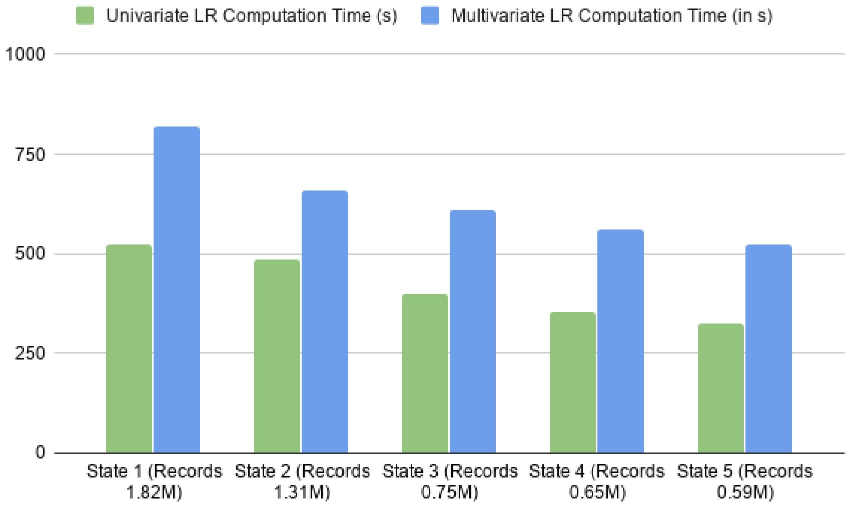


UNIVARIATE LINEAR REGRESSION

3 GOOD HEALTH AND WELL-BEING

Hypothesis Testing & Computation Time

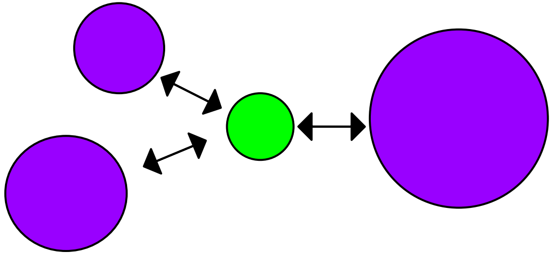
- We select the top 5 positively correlated features to then calculate p-values.
- We correct for multiple features in Multivariate Linear Regression using Bonferroni Correction.
- Due to the number of records (more than 1M for a state), we use Stochastic Gradient Descent with a learning rate of 0.1 to learn the beta coefficients.



Computation Time for 5 states

3 GOOD HEALTH AND WELL-BEING

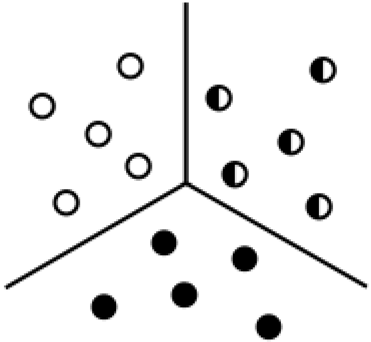
Similarity Search & Clustering



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Similarity Search

Verified



Clustering

PySpark

Benchmark District (lowest mortality)

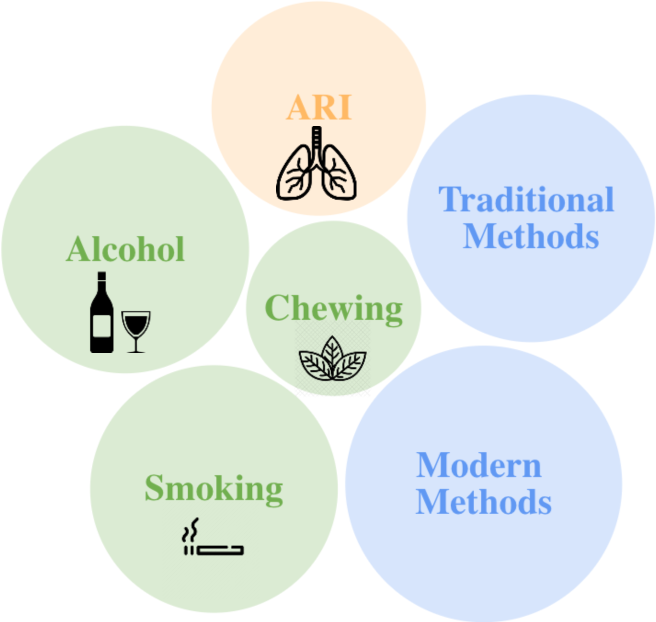
Other districts

- Similarity of each district was calculated with benchmark district.
- It was identified that districts with low literacy, no amenities indeed have high mortality.

Clustering based on chosen features verified that the clusters included similar districts; we chose number of clusters to be 4.

3 GOOD HEALTH AND WELL-BEING

Linear Regression Results


ARI

Alcohol

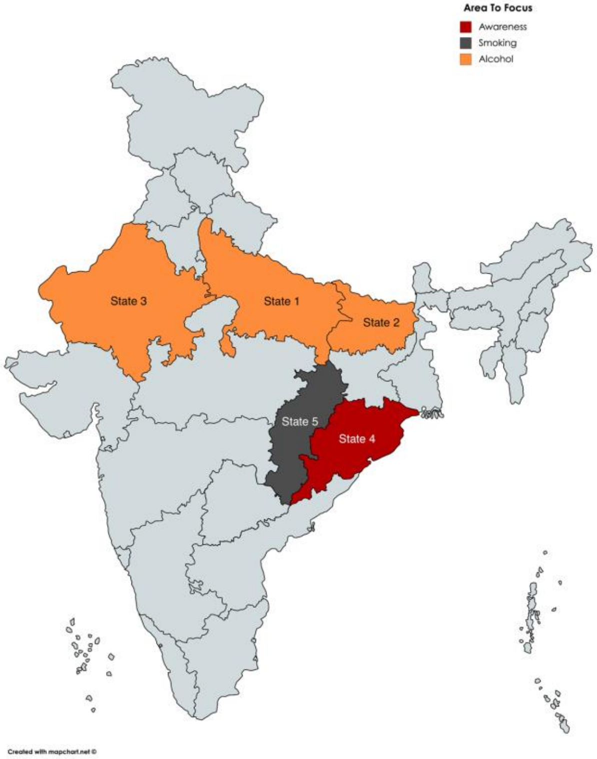
Chewing

Traditional Methods

Smoking

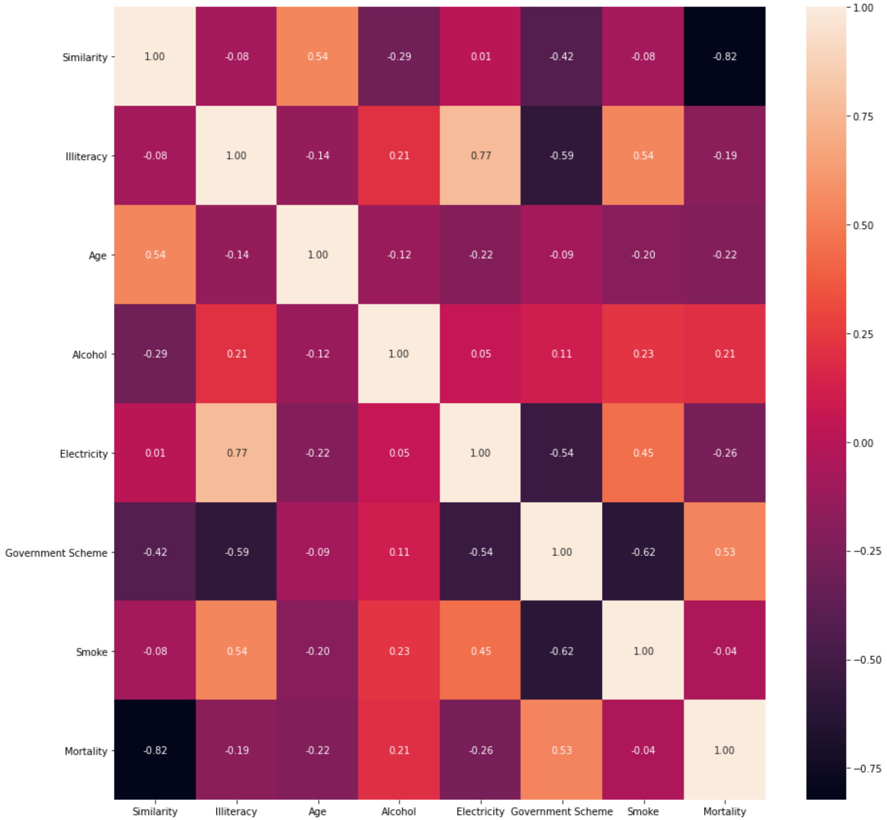
Modern Methods

Top Features That Correlated Most With Mortality



Top Feature To Work On For State (Ignoring Modern & Traditional Methods)

Similarity Search Results



Correlation between Non-biological factors and Mortality...

- Similarity (of districts with benchmark)
- illiteracy
- Age
- Alcohol
- Electricity
- Government Scheme
- Smoke

Similarity Search Results (2)

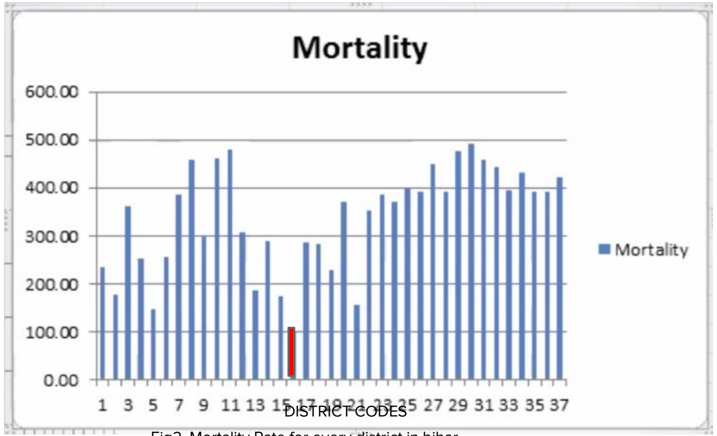


Fig2. Mortality Rate for every district in bihar

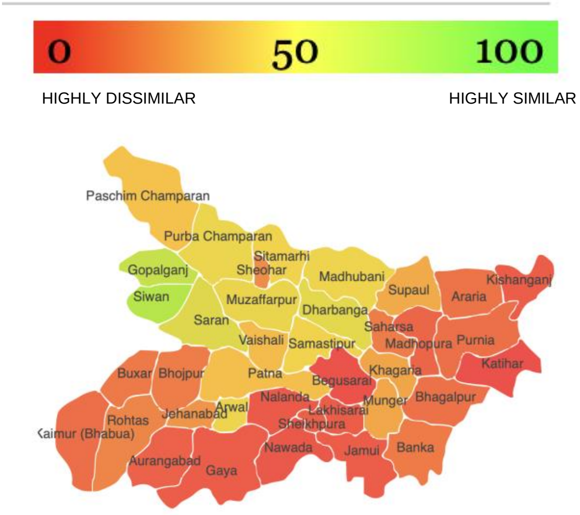


Fig1. Similarity for every district in bihar

Mortality Rate For State 2 Bihar
For Every District
Highlighted District is our Benchmark District - Siwan

Similarity For Every District with Benchmark District
Siwan For State 2 (Bihar)

3 GOOD HEALTH AND WELL-BEING

Conclusion

- Considering SDG 3 -- “Ensure healthy lives and promoting well-being for all at all ages.”, our results conclude that non biological factors also impacts IMR at some extent.

- The analysis reveals actions and awareness measures towards non biological factors such as Education, Alcohol consumption, Smoking habits etc. should be taken to reduce IMR.

*le Bonferroni's Cat