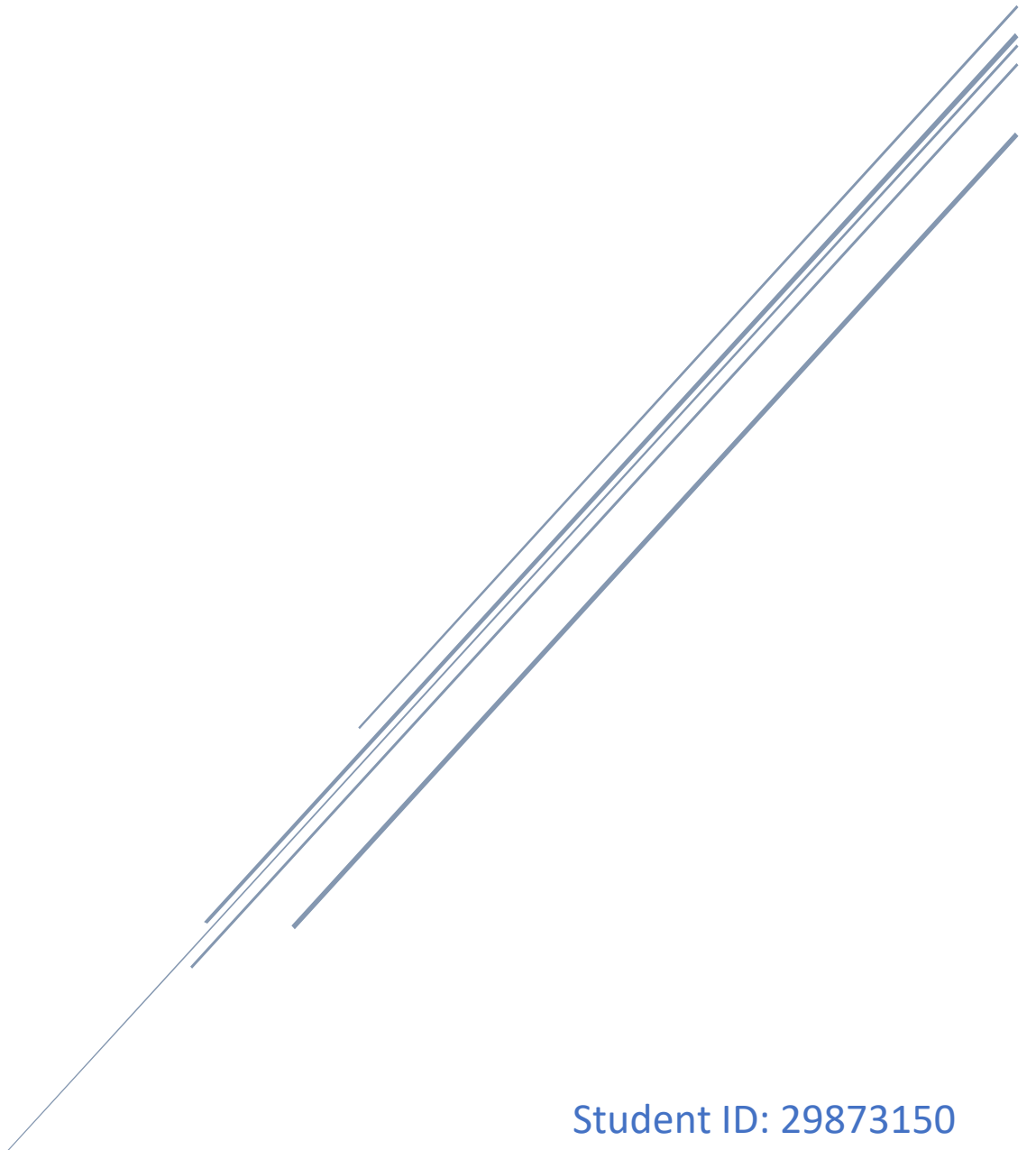


IMPROVEMENT IN GOVERNMENT POLICIES AIMED AT CLEAN INDIA MISSION THROUGH DATA SCIENCE

Prepared by MUKUL GUPTA



Student ID: 29873150
Email: mgup0003@student.monash.edu
FIT 5145 Assessment 4

Table of Contents

Project Description	3
Introduction	3
Objective	4
Data Science roles	4
Decision Modelling	4
Business Model	5
Challenges	5
Data Characteristics	6
Resources	7
Data Analysis	7
Visualisations	7
References	13

Table of Figures

Figure 1: Construction of toilets	3
Figure 2: Influence Diagram	4
Figure 3: Toilets built in villages	6
Figure 4: Percentage of villages in different states that are verified ODF	7
Figure 5: Districts with lowest percentage of toilets	8
Figure 6: States with lowest share utilized in toilet creation	8
Figure 7: Open Defecation Free villages	9
Figure 8: Open Defecation Free districts	9
Figure 9: Open Defecation Free States/UTs	10
Figure 10: ODF villages count from 2015 to 2018	102
Figure 11: States with highest ODF coverage	102

Project Description

Introduction

Data science can be used to help the decision makers of the country make better data-driven decisions. In developing countries like India, government comes out with various policies to tackle the ongoing problems within the country. With a massive population and numerous policies initiated every year, there is a need to keep a check on their functionality and progress. According to WHO, in 2012, India had 626 million people who practiced open defecation. Clean India Mission (Swachh Bharat Abhiyan) is a campaign initiated by Government of India launched on 2nd October 2014 to clean streets and cities. Clean India (Rural) objective is to achieve 100% Open Defecation Free (ODF) India by 2nd October 2019.



Figure 1: Construction of toilets

Source: <http://www.caravanmagazine.in/vantage/swachh-bharat-progress-modi-adopted-villages-nageypur-jayapur>

Lack of toilets is not only health and sanitation issue, but also dignity, climatic, privacy and economic issue. The sixth goal of sustainable development goals is 'ensure access to water and sanitation for all'. According to a survey conducted in Bahadarpur village, it was found that the major reasons for open defecation were financial constraint, waiting for government assistance, majority of time spent away from home due to labour work and habituality of going outside. (Patwa & Pandit, 2018)

Various measures are taken to spread awareness regarding hygiene and importance of constructing toilets. Individual toilets are being constructed by the government in the rural areas. Incentives are provided to Below Poverty Line households.

Objective

This project will help in evaluating the progress of Clean India scheme through data. Indian government has put down several guidelines to declare village as ODF. After the declaration of village as ODF, verification is done by the members of the sanitation/municipality team. This data is recorded corresponding to every village in India. Cost of construction of toilets are also recorded of a village. Through this data, we can find the villages in all the states that are approaching 100% ODF and which villages need more attention to construct toilets. We can also find how cost of construction of toilets relates to whether a village is ODF or not. Through this analysis, government can find how Clean India initiative is impacting the various states. If a state or village is not improving, government can take suitable actions like running more awareness campaigns or constructing extra toilets in that area.

Data Science roles

Data Engineer: Data Engineers clean and transform the dirty unstructured government data into organized data sources adhering to quality standards. Their job is to ensure that data flow smoothly from the source to destination.

Data Analyst: Data Analysts perform data visualizations to gain insights from the data. They explore the data and transform it into easy-to-understand graphical form.

Data Scientist: Data Scientists delve deep into the data and find relationships among the data elements. They evaluate why some states are still not ODF and what might be the factors behind it. They then present their analysis and findings in a user-friendly manner.

Decision Modelling

Modelling of data can be done through an influence diagram.

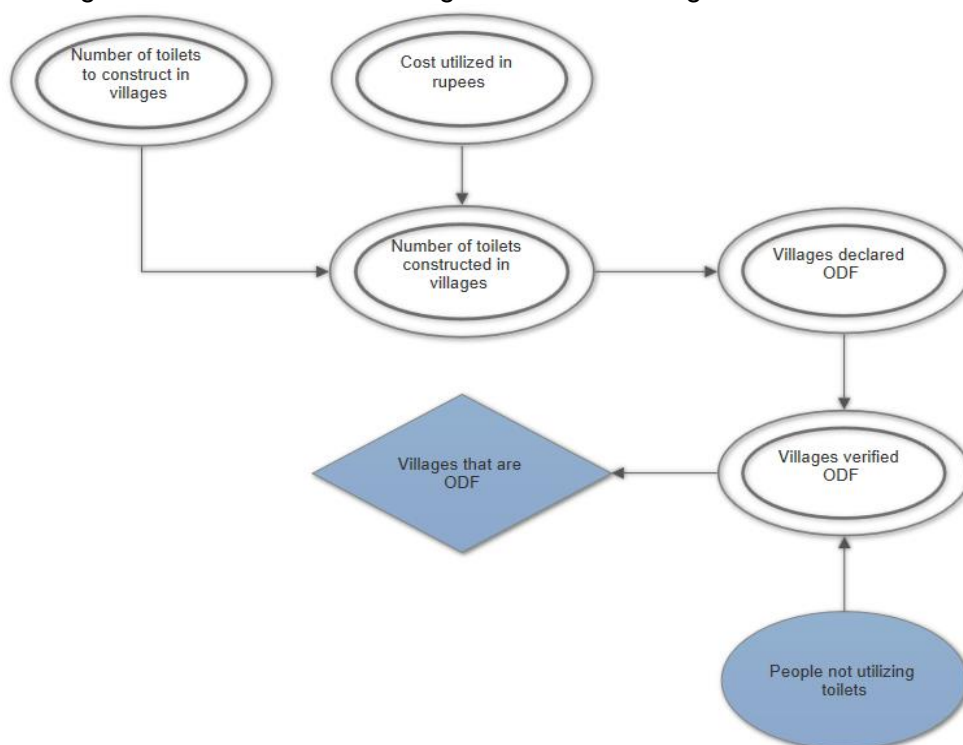


Figure 2: Influence Diagram

Built using SmartDraw (<https://www.smartdraw.com/influence-diagram/>)

Business Model

The beneficiaries of the project are:

- **Government:** Government can take various actions from the analysis and can better implement Clean India scheme. From the analysis and visualisations, it would be clear that which villages and states need more attention. This would save government's cost as they know which areas are already covered and 100% ODF.
- **Social impact:** Clean India mission in creating an impact in rural as well as in urban areas. Sanitation has favourable socio-economic returns for the people, contributing to clean environment, improving health, dignity and quality of life among other benefits. Clean India mission creates a positive impact on education, income and employment. This project would create an impact in people's lives. Villages that are becoming ODF and that are not, can easily be identified through the analyses.
- Similar projects can be implemented to evaluate other government policies. This will help to improve all government policies and in turn improve the life of every citizen in the country.

Challenges

The challenges faced would be:

Data wrangling: As there are many different sources for government data, it becomes difficult to combine them into one dataset. Also, there are many mistakes in the government data as well. Spelling of a district in one source might have different spelling in some other source. As there are different languages used in many states, it might be more difficult to combine them.

Correctness of data: Due to so many different data sources with government data, correctness of the data might be sacrificed.

False ODF declaration: There might be a case where a village is wrongly declared ODF because people are not utilizing the toilets constructed.

Awareness among people: One of the biggest challenges in the Clean India scheme is to change the way people feel about sanitation. This can be done through awareness campaigns regarding adverse health effects of open defecation. They should be taught about benefits of using a toilet and how sanitation has reduced cases of diarrhoea, vomiting and malaria among children. (Economics of sanitation, 2018)



Figure 3: Toilets built in villages

Source: <http://www.freepressjournal.in/wp-content/uploads/2017/10/>

Data Characteristics

4V's of Data and Data Processing:

1. **Data Volume:** The government data of all the villages in the country is large and dirty. Data from any different sources along with secondary research sources have to be analysed and wrangled to produce clean data which can be analysed. This means data in terms of GBs would have to be cleaned. This is a one-time process though. We also collect data from all the villages in the states on a regular basis to get the updated version. This data would also be huge.
2. **Data Velocity:** During the clean India movement, toilets are built on a regular basis on a regular basis in the needed villages. This data is updated on a regular basis (every minute) in order to ensure regular up-to-date visualisations.
3. **Data Variety:** Toilet condition with images, cost of construction, number of toilets built are collected.
4. **Data Veracity:** Standards are maintained to declare and verify a village as ODF. Data is prone to human errors. Some villages might be falsely declared and verified as ODF.

Lifecycle: Continuous reports and visualisations should be maintained till each village in every state is identified and verified as ODF. The reports should be shared with public so that everyone knows how the government policies are doing in the long run.

Data Processing: Primary and secondary data which is huge needs to be analysed using big data technologies like Hadoop and Hive. Stream-processing would not be viable as the data is not changing every few seconds. Visualisation technologies like Tableau, R or Python can be used for report generation. A mobile application can be used by various government sanitation individuals who declare and verify a village as ODF. They can update the status of a village as ODF through the application. RDBMS or MongoDB should be used to store data coming from various villages and states.

Resources

Data for the analysis is extracted from government sites and data stores (government data), primary and secondary research data, sanitation organization data (approved by the government) that would declare and verify a village as ODF.

This project analyses the Clean India mission data available on: https://quip.com/-/blob/ABOAAAAynyX/Vx3EaNEYYH6LZ3VfDviMRA?s=MI7sAb0GRLK6&name=SBM_DummyData.csv

Data for the project can be taken from government sources like:

<https://data.gov.in/catalog/daily-data-open-defecation-free-odf-under-swachh-bharat-mission>

Data Analysis

Hadoop is an open-source distributed data processing platform that can be used for big data analytics. It partitions the data and sends each partition across different nodes in the cluster. Each node performs similar operations on different data. At the end, partitioned solutions are integrated to get the result. Hive can be used to query on top of Hadoop for data analysis. We can use Tableau or R for visualisation and reports. Dashboard can be generated from the visualisations as well.

Visualisations

1. Villages are grouped as states and ODF percentage is found in each state. It was found that states like A and N Islands, Arunachal Pradesh, Tripura require the most attention. Also, states like Sikkim, Puducherry are 100% ODF.

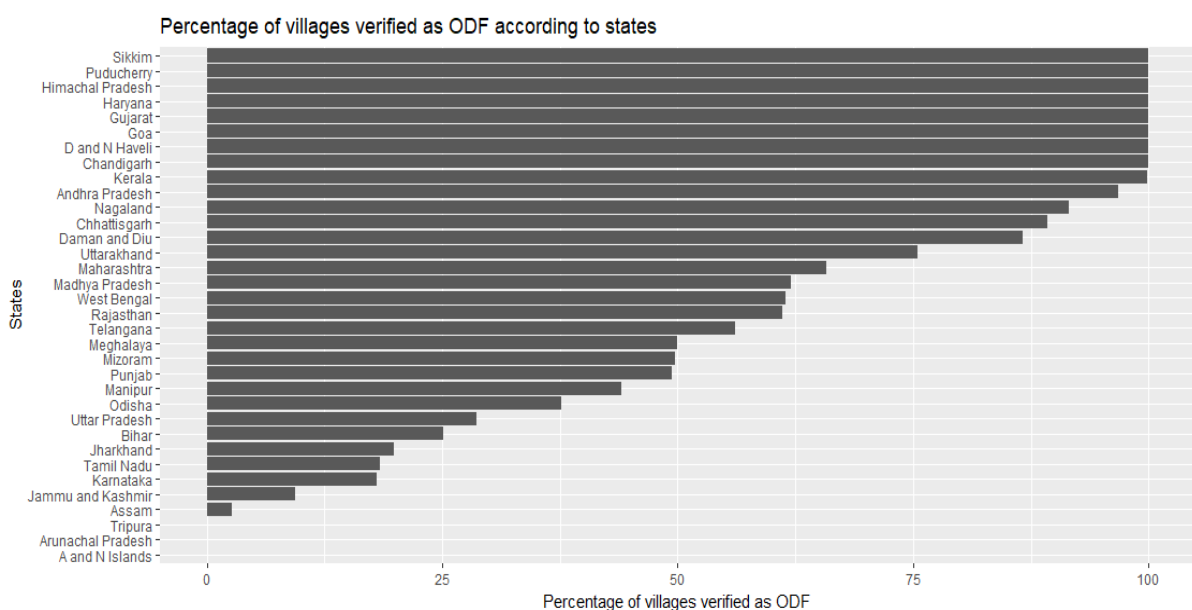


Figure 4: Percentage of villages in different states that are verified ODF

- Villages are grouped as districts and ODF percentage is found in each district. Districts with lowest ODF percentage are displayed. It was found that districts like Nicobars, Pondicherry, Madhubani require the most attention.

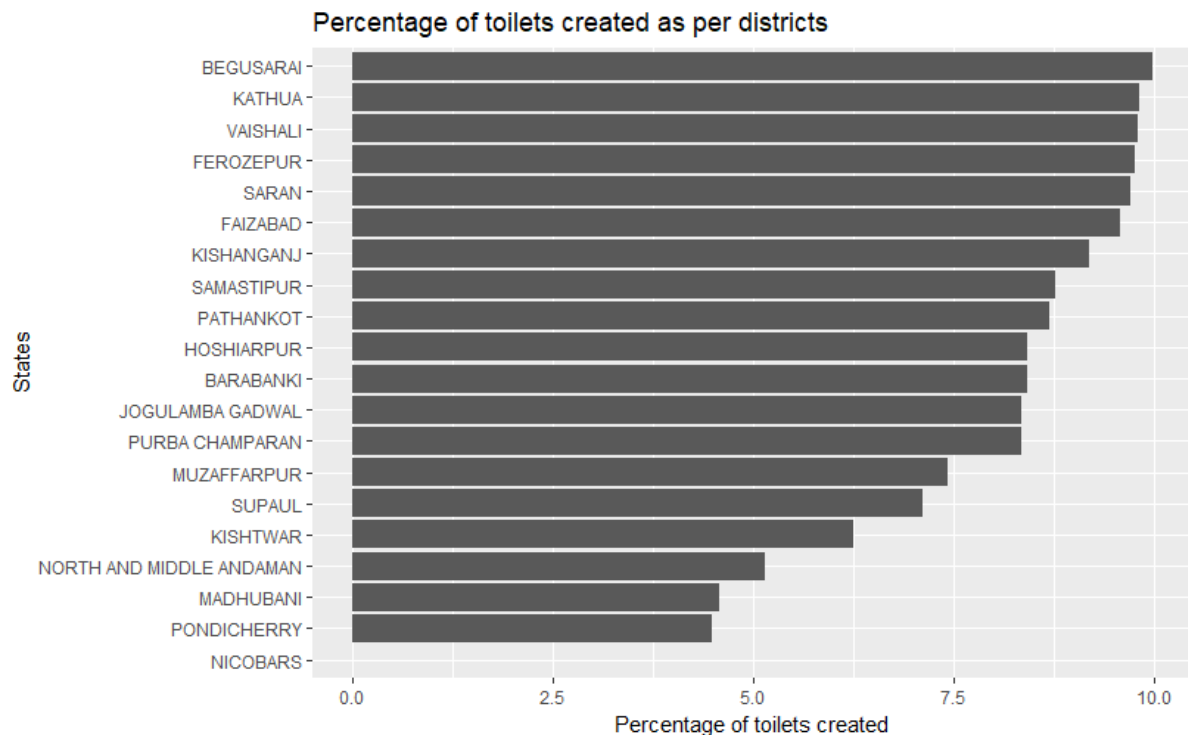


Figure 5: Districts with lowest percentage of toilets

- We visualise states with their respective toilet percentage. Each bar is filled with total cost that is utilised per toilet in each state.

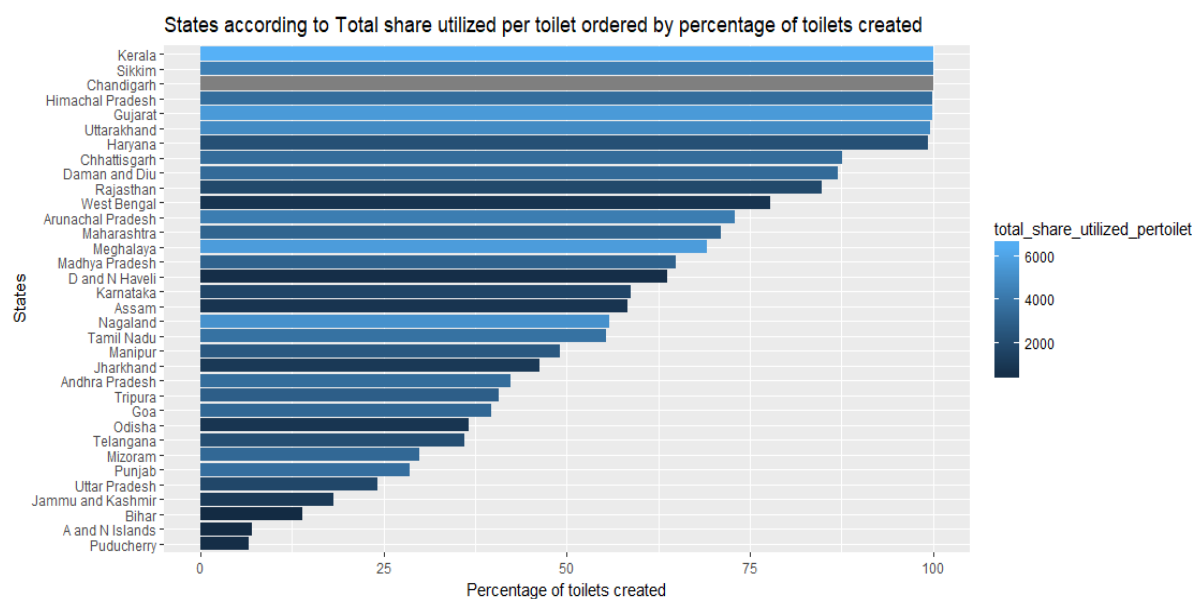


Figure 6: States with lowest share utilized in toilet creation

States with lowest percentage of toilets created have lowest share utilized per toilet. This means that these require more share from the government as compared to states like Kerala and Sikkim where 100% of the toilets are created.

Government can also look into other factors which have led to such low percentage of toilets created in these regions.

Other visualisations that can be implemented are:

4. How Clean India initiative is helping India in becoming Open Defecation Free (ODF).

This graph shows 5,12,991 villages are ODF on 14 October 2018.

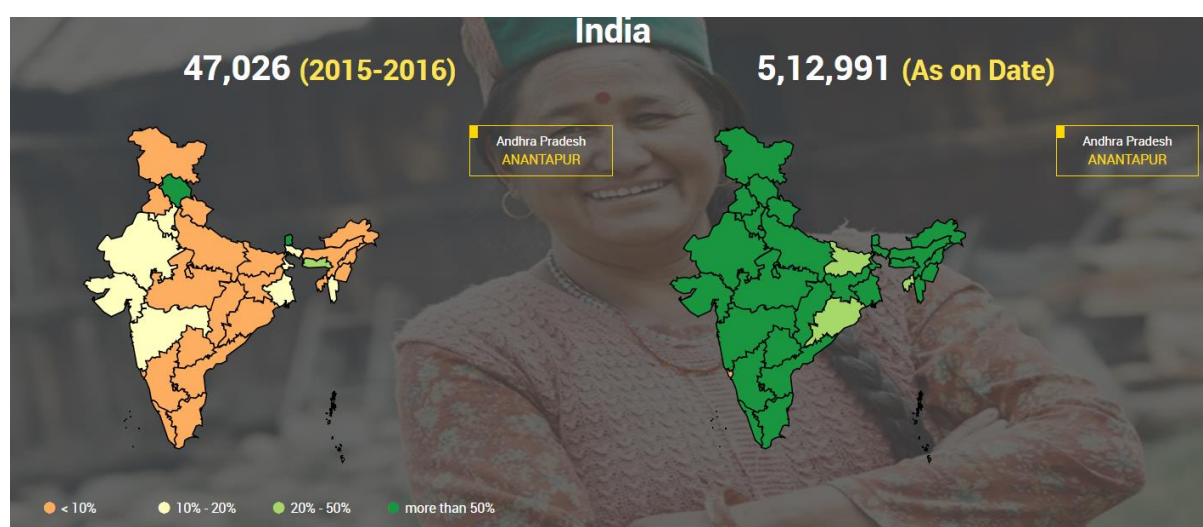


Figure 7: Open Defecation Free villages. (on 14 October 2018)

Source: <https://sbm.gov.in/sbmdashboard/Default.aspx>

Following graph visualises the ODF districts on 14 October 2018.

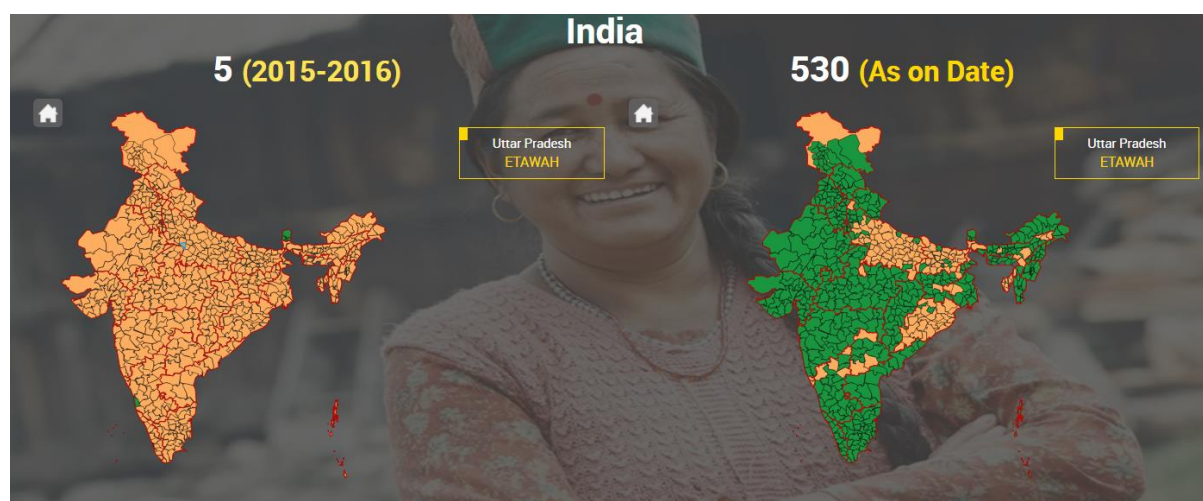


Figure 8: Open Defecation Free districts. (on 14 October 2018)

Source: <https://sbm.gov.in/sbmdashboard/Default.aspx>

Following graph specifies the states and Union territories of India that are ODF. 25 states/UTs are ODF on 14 October 2018.

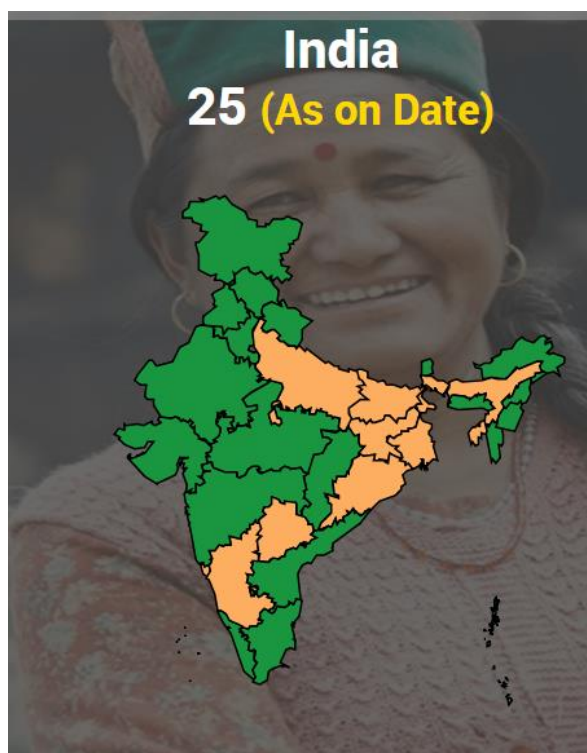


Figure 9: Open Defecation Free Staes/UTs. (on 14 October 2018)
Source: <https://sbm.gov.in/sbmdashboard/Default.aspx>

5. ODF villages details at a glance.

Total of 3,78,094 villages are verified ODF of the 5,12,991 villages that are declared ODF. ODF percentage of India is **85.22%** on 14 October 2018.

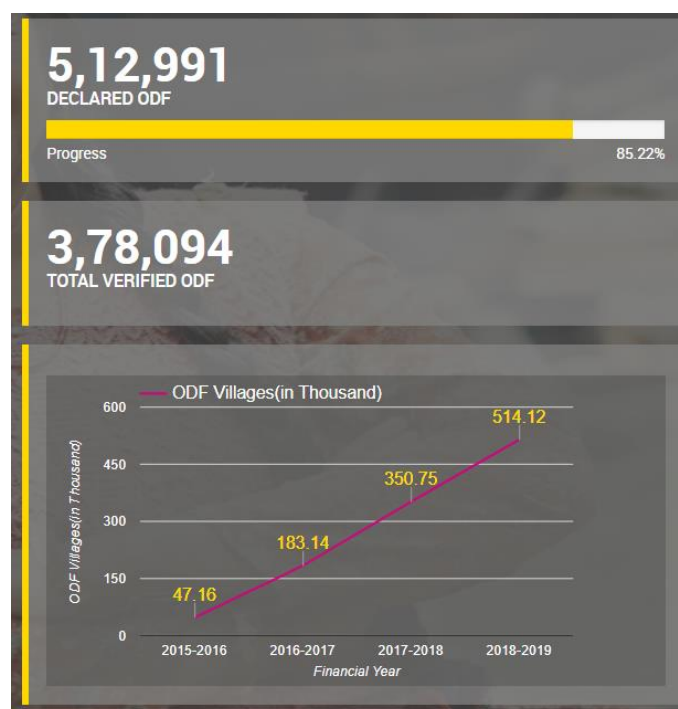


Figure 10: ODF villages count from 2015 to 2018
Source: <http://sbm.gov.in/sbmdashboard/ODF.aspx>

6. States which are still struggling to be ODF along with states which are 100% ODF.

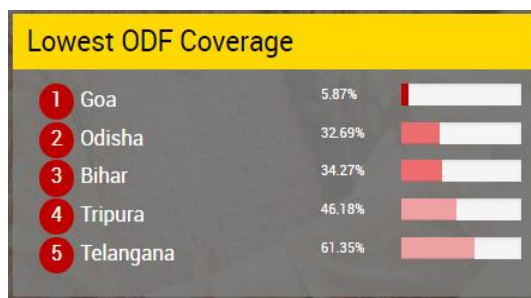


Figure 11: States with lowest ODF coverage
Source: <http://sbm.gov.in/sbmdashboard/ODF.aspx>

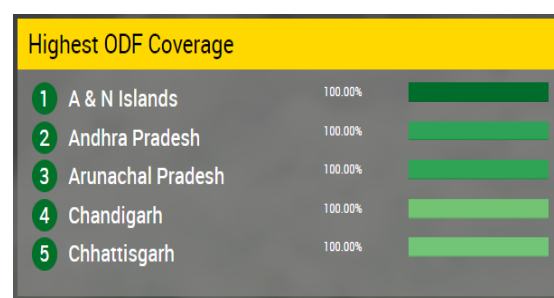


Figure 12: States with highest ODF coverage
Source: <http://sbm.gov.in/sbmdashboard/ODF.aspx>

These visualisations not only help the decision makers of the country but also helps in motivating every individual in the country. Everyone can see how India is progressing day-by-day in their journey towards cleanliness.

Other ways in which technology can help achieve “Clean India”

Improving efficiency of Garbage cleaners

Problem: The public worker system in India lack “incentives for good performance”.

How tech can help: Citizens in a locality can rate the streets based on their cleanliness and hence incentivize the cleaning personnel in that area. Scores can be collected over a period to identify the cleanest streets around the country/area. Best Cleaning personnel award should be awarded to the ones with top consistent performances. This would help improve morale as well as increase the attendance rate of the cleaning personnel. (Garg, 2014)

Monitoring of garbage

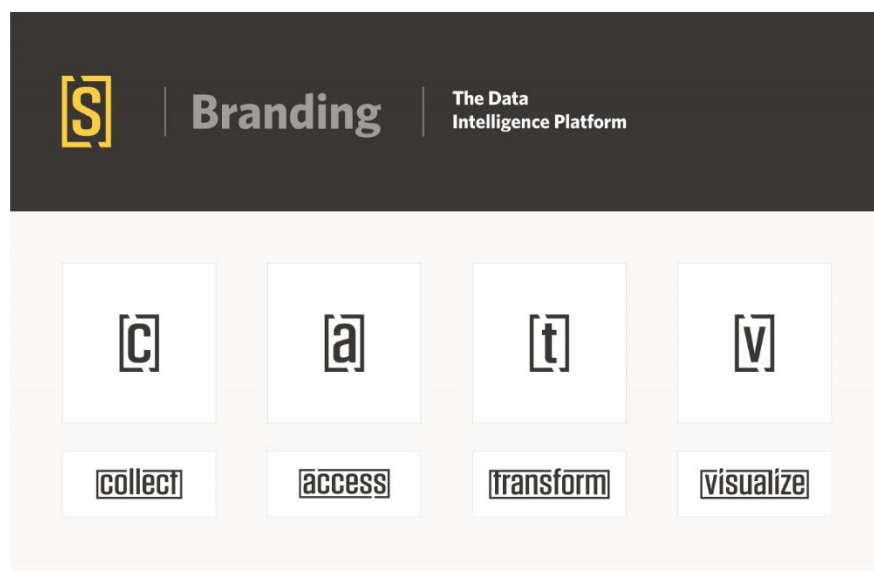
Problem: About 60,000 tonnes of garbage is generated every day in India. There is a problem for treatment and disposal of such huge amount of garbage with the Municipal Corporations. In India, garbage collection is outsourced to third party agencies. The processing includes generation, segregation, storage, collection and transportation, processing and disposal of waste. (Garg, 2014)

How tech can help: We can monitor and supervise the parameters by asking the citizens in the region to rate the cleanliness levels every day. This can then be directly mapped to the efficiency of the agencies and would help take the corrective steps.

SOCIALCOPS: One of the companies that is helping the government take better data-driven decisions



SocialCops is a data intelligence company based in New Delhi, India that is tackling helping organisations like United Nations, the World Bank, Unilever and others make better decisions through their platform. They are helping governments, NGOs, philanthropists tackle the most important issues through data. Their platform includes:



- **Collect**
This mobile data collection tool helps to create surveys to collect data from remote corners of the world on an Android app.
- **Access**
This external data engine extracts, cleans, verifies the reliable data from public and government sources.
- **Transform**
This transformation engine converts the unclean and dirty data into clear insights.
- **Visualize**
This is the analytics dashboard helps to take the important decisions by revealing the insights from the data.

Conclusion

Data science and data analytics is becoming popular among the policymakers around the world leading to better and more efficient public policies. With more and more government data becoming accessible every day through government sites, various stakeholders approach, analyse or use government's data for various purposes (Numanović, 2017). With data science techniques being widely used in various scientific fields of medicine and astrophysics, applying data science for social good and well-being is also necessary.

References

- Economics of sanitation. (2018, Sep 03). The Assam Tribune Retrieved from <https://search-proquest-com.ezproxy.lib.monash.edu.au/docview/2098628243?accountid=12528>
- Garg, V. (2014, November 15). Retrieved from Quora: <https://www.quora.com/What-are-some-ways-we-can-use-technology-to-help-achieve-Clean-India>
- Numanović, A. (2017, July 12). Retrieved from Medium: <https://medium.com/@numanovicamar/https-medium-com-numanovicamar-data-science-the-next-frontier-for-data-driven-policy-making-8abe98159748>
- Patwa, J., & Pandit, N. (2018). Open defecation-free India by 2019: How villages are progressing?. Indian Journal of Community Medicine, 43(3), 246. Retrieved from <http://link.galegroup.com.ezproxy.lib.monash.edu.au/apps/doc/A555348739/AONE?u=monash&sid=AONE&xid=17bb23f3>
- Sujoy Mojumdar, D.-S. (2014). *Guidelines for Swachh Bharat Mission (Gramin)*.
- WHO/UNICEF. (2012, March). http://www.who.int/water_sanitation_health/monitoring/jmp2012/fast_facts/en/. Retrieved from WHO website.