# Respiratory Ailments: Air Pollution & other risk factors

**By Group - 1:**

- Armeet Singh Luthra (200185)
- Shivanie (221020)
- Jahnabi Kachari (231080046)
- Aditya V (220082)

## Acknowledgement

We, the students of group - 1 would like to express our profound gratitude towards **Dr. Dootika Vats**, our academic and project instructor for MTH208A (Data Science Lab I), for her guidance and constant supervision throughout the process and providing creative ideas and necessary information regarding the project which led to the completion of this project.

It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course lecture.

## Introduction

### Understanding the possible influential factors of Acute Respiratory Infections (ARI)

Acute Respiratory Infections (ARI) encompass a broad spectrum of respiratory illnesses that affect the upper and lower respiratory tracts. ARIs are a leading cause of morbidity and mortality in the country. They contribute substantially to the global burden of disease, particularly in vulnerable populations such as children, the elderly and the individuals with compromised immune systems. These infections are a significant health concern due to their potential for severe complications and widespread prevalence. The consequences of ARIs extend beyond individual health, influencing communities, healthcare systems and economies. This project aims to delve into the factors that contribute to the occurrence of ARIs.

Understanding the interplay between environmental conditions, host factors, preventive measures and broader societal factors is crucial for developing effective strategies to mitigate the impact of ARIs.

## Motivation

- IMPACT ON VULNERABLE POPULATIONS:

Children, pregnant women, the elderly and individuals with underlying health conditions are more susceptible to severe outcomes from ARIs.

- ROLE OF HOUSEHOLD VARIABLES

1. Tobacco usage

Tobacco smoke contains a number of harmful chemicals. Both active smoking and exposure to secondhand smoke can have harmful effects on the respiratory system, making individuals more susceptible to respiratory tract infections.

2. Usage of dirty and rancid cooking oil

Dirty cooking oil may contain harmful substances such as free radicals, carcinogens and toxins produced during the breakdown of oil during cooking. The ingestion of these harmful substances leads to systemic inflammation and compromises the body's immune response.

3. Literacy rates

The level of education and awareness within the household regarding respiratory health, recognizing early symptoms and observing preventive measures of ARIs largely influences the family's ability to take appropriate actions to reduce the spread of infections.

- AIR POLLUTION

Air pollution, especially high levels of particulate matter (PM) such as PM2.5 and PM10 can irritate the respiratory system. Particulate matter penetrates deep into the lungs, causing inflammation, and irritation of the airways. Not to neglect the fact that prolonged exposure to air pollutants can weaken the immune system's ability.

## Data Description

### DATA:

To analyse the contribution of possible causal factors of ARIs we have collected the primary data from the Open Government Data Portal designed, developed and hosted by the National Informatics Centre (NIC), a premier ICT organisation of the Government of India under the aegis of the Ministry of Electronics & Information Technology. Datasets were also scrapped from Indiastat, owned by Datanet India Pvt. Ltd. All the datasets obtained were related to the surveys done in 2011-12.

## OBTAINING THE DATA:

### The data was obtained from the following webpages:

Links to the Datasets →

1. National Family Health Survey(NFHS):

https://data.gov.in/resource/all-india-level-and-state-wise-key-indicators-nfhs-3-and-nfhs-4

2. City wise Ambient Air Quality Quality:

https://data.gov.in/resource/city-wise-ambient-air-quality-year-2011#api

3. State wise ARI cases and deaths:

https://www.indiastat.com/table/health/state-wise-number-cases-deaths-due-acute-respirato/707964

4. PM 2.5 levels:

https://urbanemissions.info/india-air-quality/india-satpm25/

National Family Health Survey(1) and the City wise Air quality(2) datasets were scraped using API. First we had to register on the data.gov.in portal to be able to generate an API key. Using this API key appropriately in the URL, by following the steps given on the portal itself, we were able to obtain the webpage link which contained the datasets in an XML code. Libraries like tidyverse ,dplyr and XML were used in Rstudio to send a GET request and extract XML content as a character string, which was processed in next steps for analysis.

State wise ARI cases and deaths(3) dataset was downloaded as an excel file from the webpage.

PM 2.5 levels were obtained as a .csv file.

XML content in the character string saved in Rstudio, was then parsed followed by extracting specific nodes and looping through the extracted nodes to populate the vectors which represented the columns of the dataset. A data frame was created from the extracted vectors, which formed our dataset. This dataset was converted to a tibble format for further analysis. The datasets were cleaned by removing the NULL rows and properly renaming the states so that merging of the datasets happens smoothly.

Only those columns in the NFHS dataset were retained, which might have been affecting the % prevalence of ARI. Since the data was for NFHS-3 and NFHS-4, average of the values in the two datasets was computed and used for further analysis.

From the city wise ambient air quality data, state wise air quality data was calculated as an average of the pollutant levels of the cities which lie within a state.

The state wise ARI cases and deaths dataset, which was obtained as an excel file, was converted to a .csv file for further analysis

The PM2.5 levels .csv file was modified to retain the 2011 state wise pollution data.

### Challenges faced while obtaining the data:

- Finding a good and reliable source of data related to Respiratory illness specific to a particular year, was tough. National level Government surveys are authentic but they are undertaken only after a fixed number of years. After an elaborate search, year 2011 was finalised. Conflict issues were resolved by restricting data to sources stemming from the 2011 census, and restricting data for factors like domestic variables to a single credible survey to set up justifiable correlations.
- Some data were obtained using web scraping on websites. In those cases, the parsing of data was a challenge. Even after obtaining and parsing the data, the data had to be cleaned and put into a specific format to make it suitable for our use case.While cleaning, we had to assign data types to values in proper formats like integer, string, etc. All these type casting was done manually and had to be assigned after proper review of the data.
- Finding data that go well with each other and can be related to one another without arising conflicts, in terms of data from varying surveys on different population samples, different states/union territories, different surveys, was difficult. For example in one dataset name of the state Chhattisgarh was written as Chattisgarh. This was leading to loss of data while merging the datasets. State names had to be checked manually and made consistent in all the datasets.

## RELEVANT FEATURES OF DATASETS

### Deaths_2011.csv

| Variable name | Description |
|---|---|
| State | State/UT name |
| Male.Cases | Male ARI case count |
| State | State/UT name |
| Male.Deaths | Male Death count due to ARI |
| Female.Cases | Female ARI case count |
| Female.Deaths | Female Death count due to ARI |
| Total.Cases | Total ARI related cases |
| Total.Deaths | Total ARI related deaths |

| Variable name | Description |
|---|---|
| populations | State Population |
| Cases/population | Ratio of Total cases to State population |
| Total Deaths/Cases | Ratio of Total deaths to Total cases of ARI |
| Male Deaths/Cases | Male death to Male ARI cases ratio |
| Female Deaths/Cases | Female death to Female ARI cases ratio |

This dataset has 34 observations.

### PM25.csv

| Variable name | Description |
|---|---|
| State | State/UT |
| PM2.5_ug_m3 | Annual average of PM 2.5 levels |

This dataset has 36 observations.

### State_wise_ARI_and_factors.csv

| Variable name | Description |
|---|---|
| State | State/UT name |
| Households_using_clean_fuel_for_cooking | % of households which use clean fuel for cooking |
| Women_who_are_literate | % of women(15-49years) who were literate |
| Men_who_are_literate | % of men(15-49years) who were literate |
| Prevalence_of_ARI_under_5_years | % prevalence of ARI in children under the age of 5 years |

| Variable name | Description |
|---|---|
| Tobacco_use_women | % of men(15-49years) consuming tobacco in any form |
| Tobacco_use_men | % of women(15-49years) consuming tobacco in any form |

This dataset has 37 observations.

**air_quality.csv**

| Variable name | Description |
|---|---|
| State | State/UT name |
| City | Name of city |
| SO2_Annual_Average_g_m3 | Citywise Annual average of sulphur dioxide levels |
| NO2_Annual_Average_g_m3 | Citywise Annual average of Nitrogen dioxide levels |
| Air_Quality_of_NO2 | Qualitative description of NO2 levels |
| PM10_Annual_Average_g_m3 | Citywise Annual average of PM 10 levels |
| Air_Quality_of_PM10 | Qualitative description of PM10 levels |

This dataset has 182 observations.

**statewise_pollution.csv**

| Variable name | Description |
|---|---|
| State | State/UT name |
| SO2_Annual_Average_g_m3 | Average sulphur dioxide levels |
| NO2_Annual_Average_g_m3 | Average nitrogen dioxide levels |
| PM10_Annual_Average_g_m3 | Average PM 10 levels |

This dataset has 28 observations.

### Possible biases/flaws in the DATA

Data related to prevalence of ARI might vary depending on percentage of ARIs actually reported in the survey (children)/ recorded officially (general) which may vary from state to state leading to possible inaccuracies in data as well as inferences drawn.

Domestic factors like tobacco consumption might are also subject to false reports from households, especially in states where tobacco smoking is more frowned upon.

POSSIBLE BIASES IN THE DATA

INTERESTING QUESTIONS TO ASK FROM THE DATA

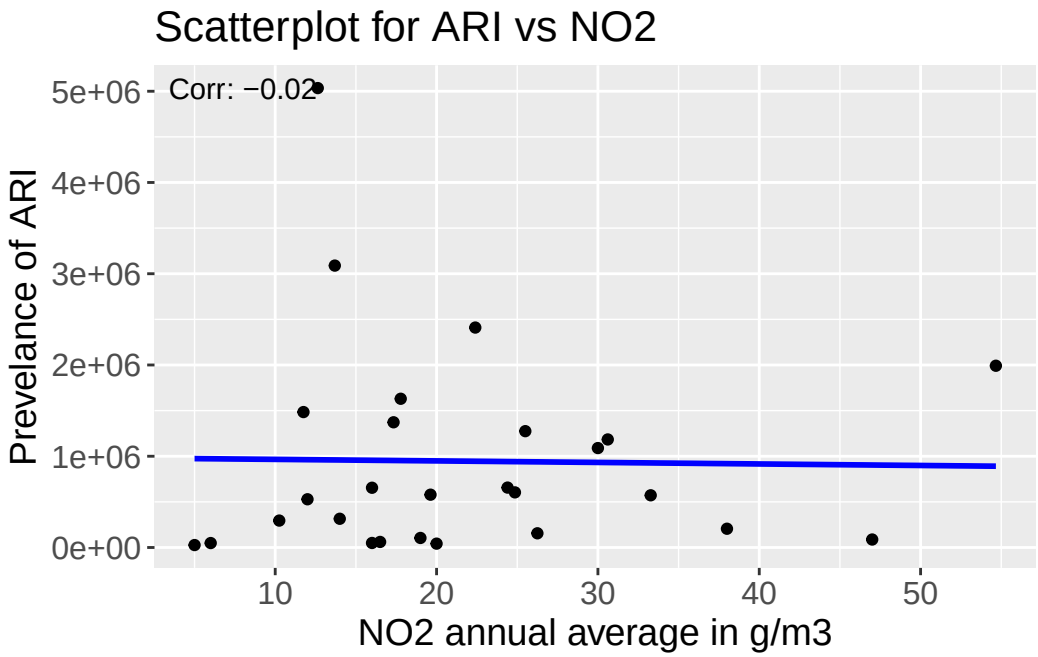
As we explore the relationship between respiratory illness and various risk factors, our analysis aims to answer questions such as:

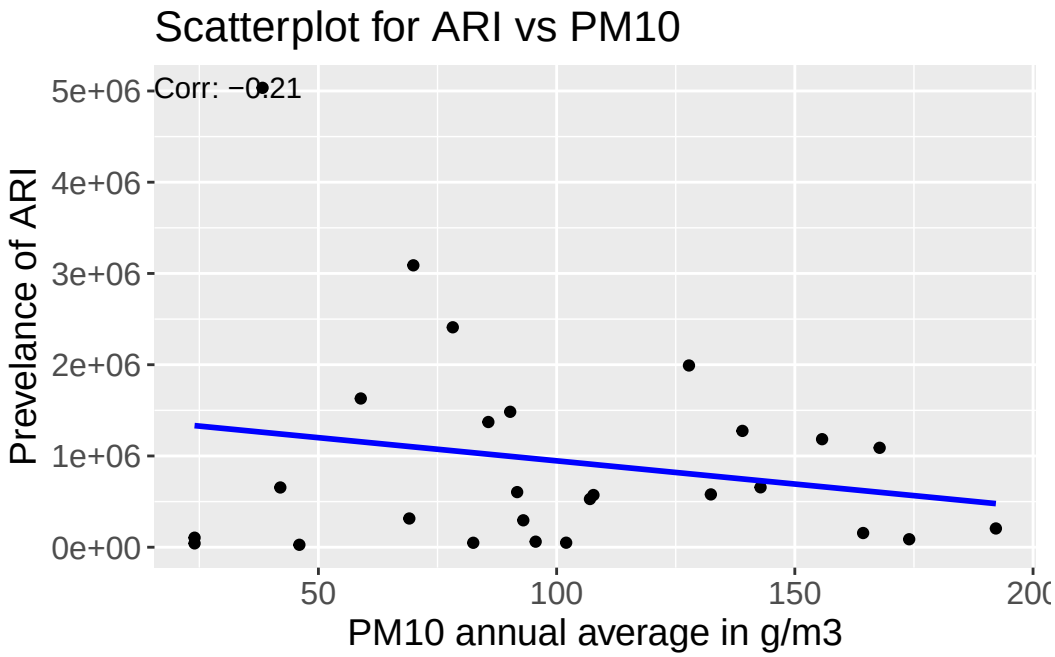
1. How does air pollution (NO2, SO2, PM2.5, PM10) correlate with ARI prevalence in children?
2. Which states do the best and worst in mitigating ARIs ?
3. Which states are the most polluted?
4. How do household variables like clean cooking fuel usage and tobacco use impact respiratory health?

## Important visualizations

The most important question that we hoped to gain insights into over the course of this project was which aspects of pollution could be most strongly linked with incidences of ARIs.

The first way we explored this was to see the correlation between cases per state and compare it with various pollutants.



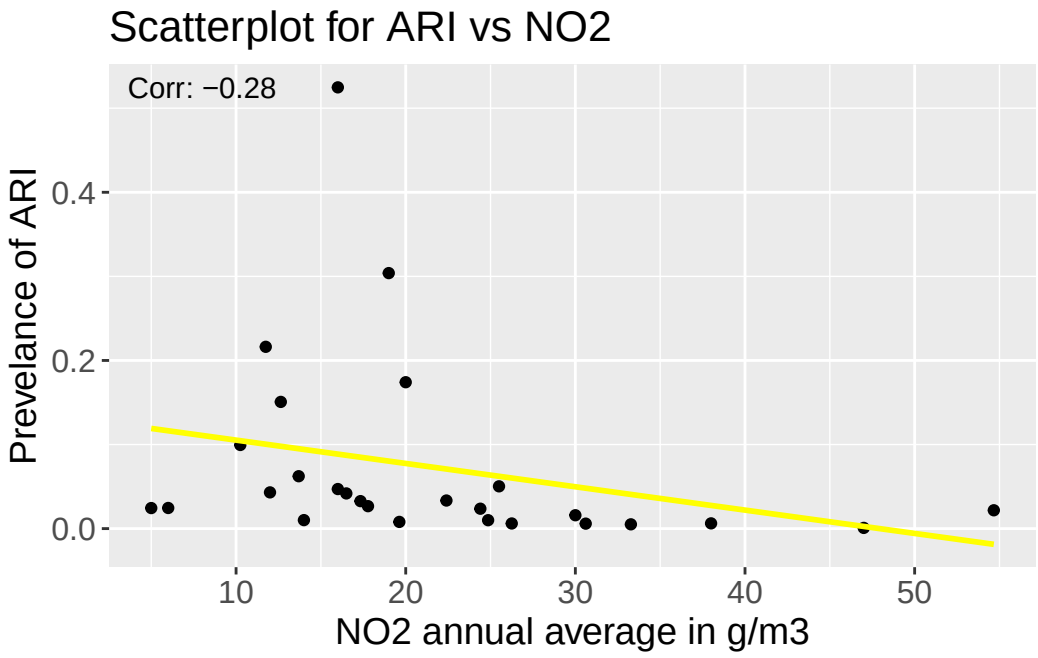



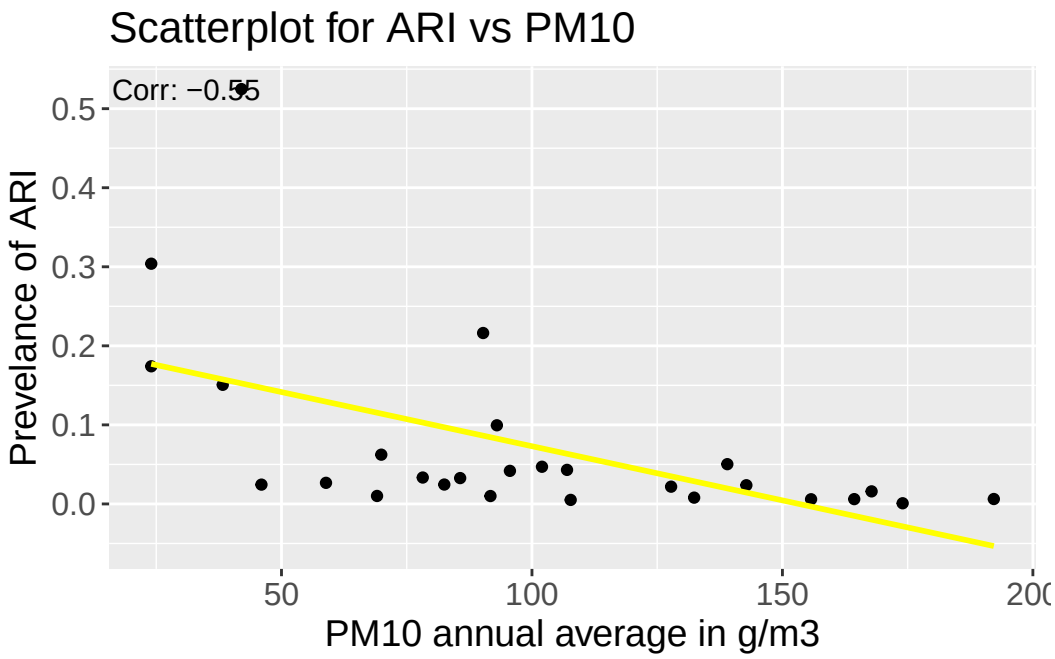

The plots clearly shows strange tendencies and provides no meaningful insights.

However an obvious flaw with this approach is the fact that different states have different populations, so a more meaningful metric would be to see prevalence of ARI and compare it

with various pollutants. So we used the census data to obtain state populations in 2011 and found out the number of cases of ARIs as a ratio of state population

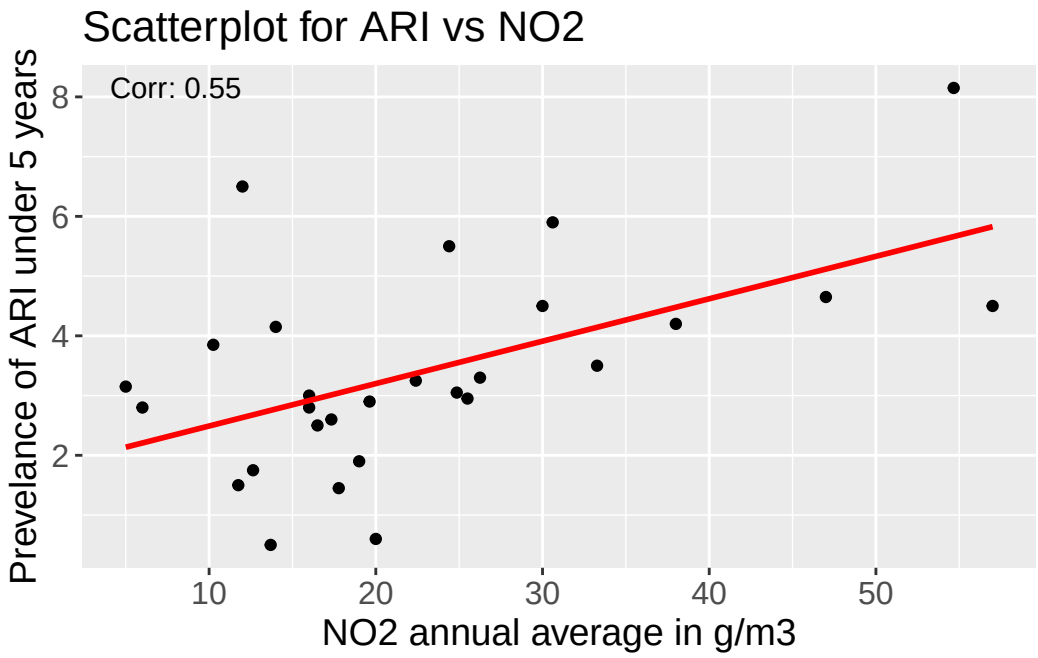
.

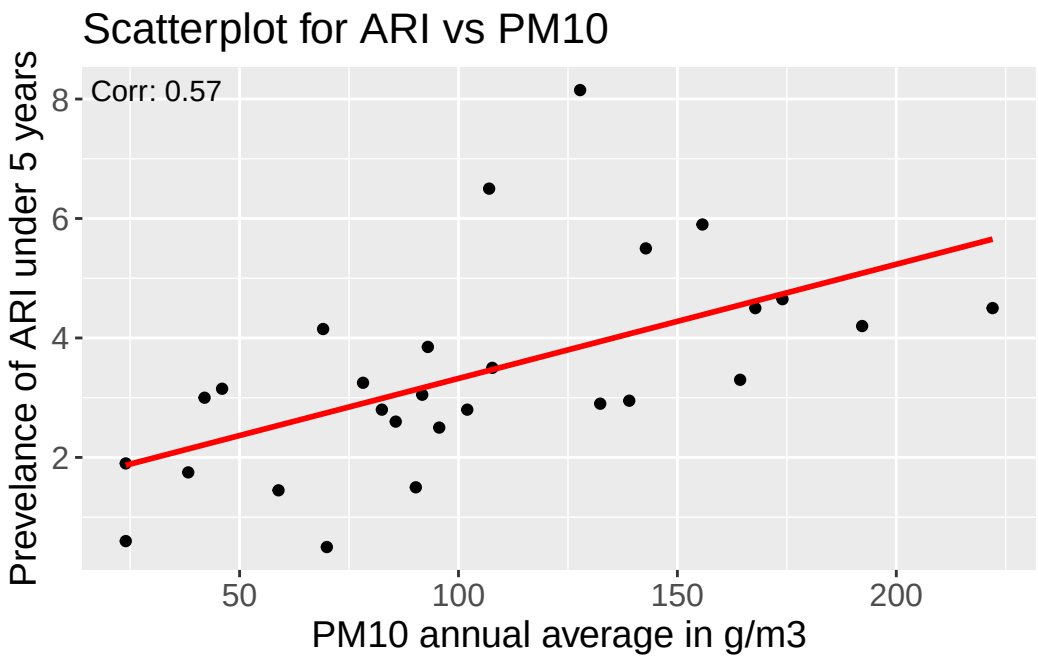


Strangely and unintuitively, even this plot seems to suggest there is no direct correlation between prevalence of ARI in a population and the effect of pollutants.

However upon doing a similar analysis of prevalence of ARIs in children aged below 5 years against various pollutants, we found that there was a much more clear correlation between the two.

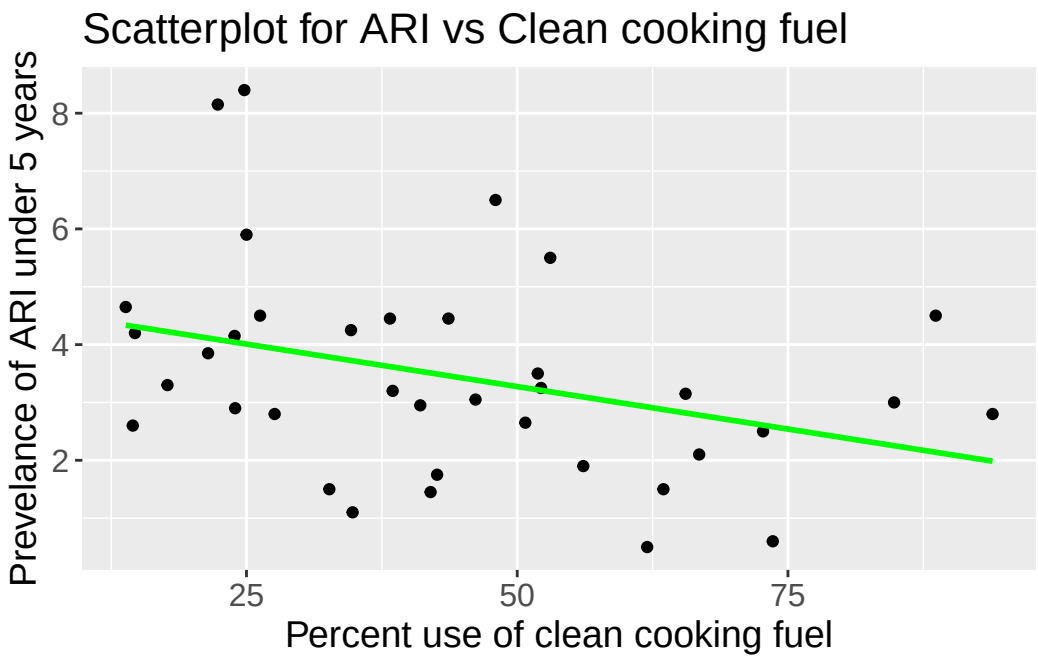




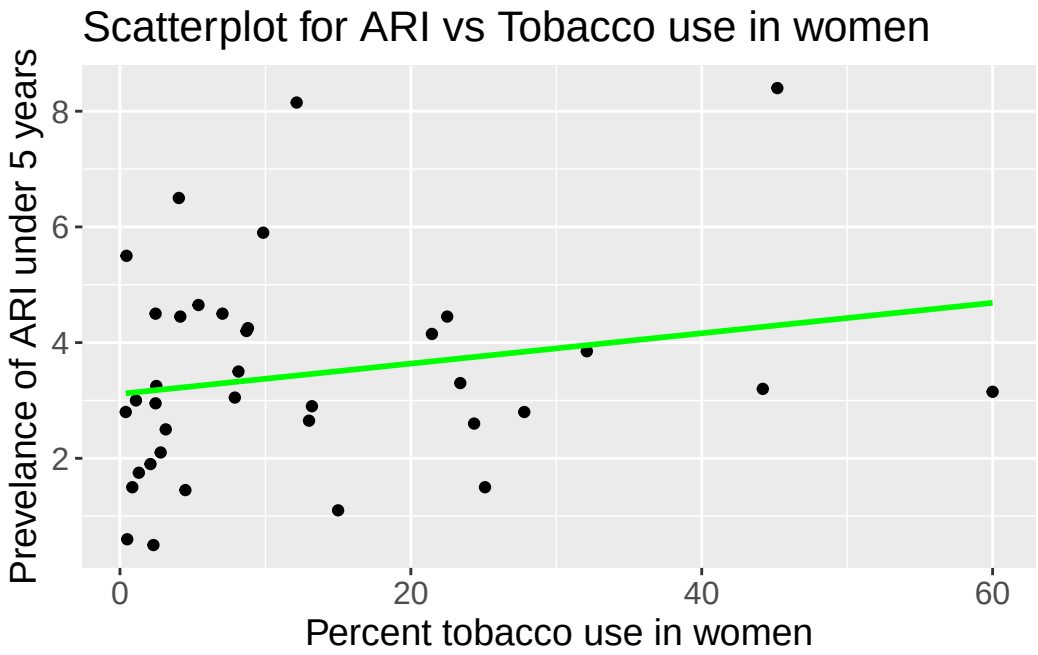

A possible insight that we might draw from this is that the affect of a pollutant on children is significantly larger than its effect on the general population (which has a larger proportion of adults). Thus, the chances of an adult getting affected by an ARI is probably more equally influenced by other factors.

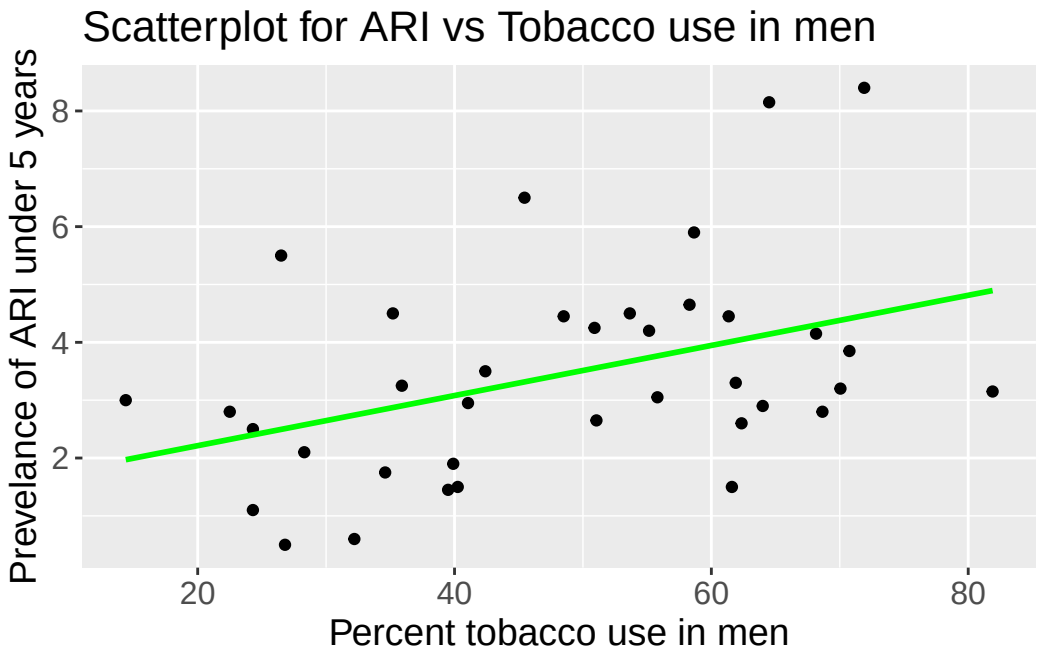
We can also see from the plots that NO2 and PM10 particles are more closely linked to the risk of ARIs than SO2 and PM2.5. This agrees with the knowledge of NO2 being a much more poisonous gas than SO2.

Following this, we did an analysis of the impact of other domestic factors such as tobacco use, literacy in households, and usage of clean cooking fuels on the risk of ARIs.




The impact of tobacco usage had the highest impact. Another interesting inference was the impact of male tobacco smoking vs female tobacco smoking in households. There is a much sharper correlation in the case of male smoking. This raises interesting questions about the forms of smoking preferred by men versus women, as well as the amount of time smoked by each sex.

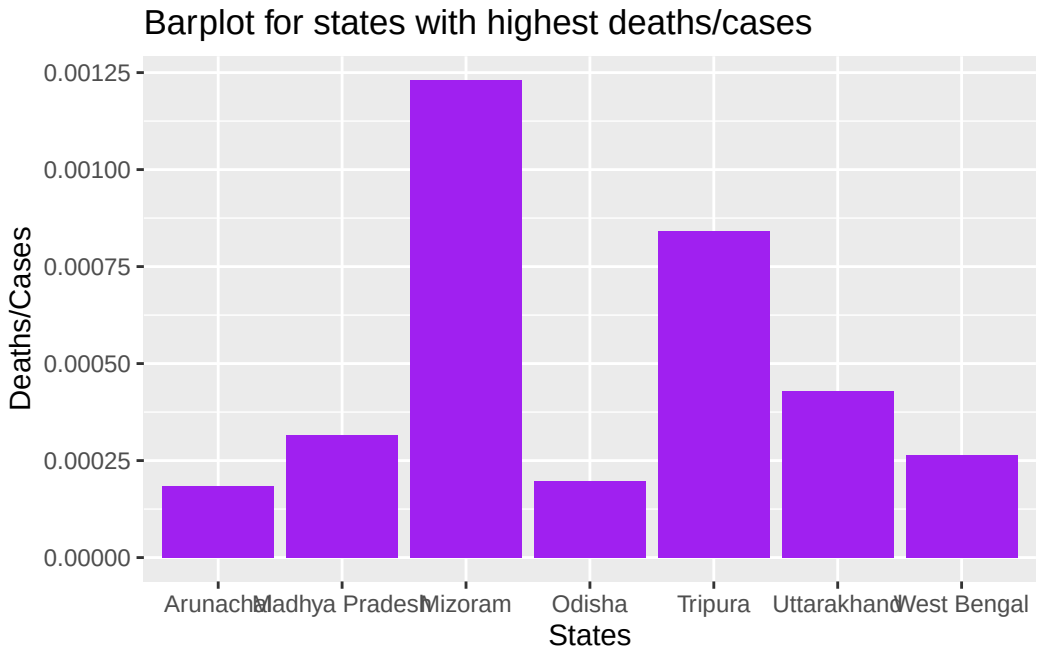


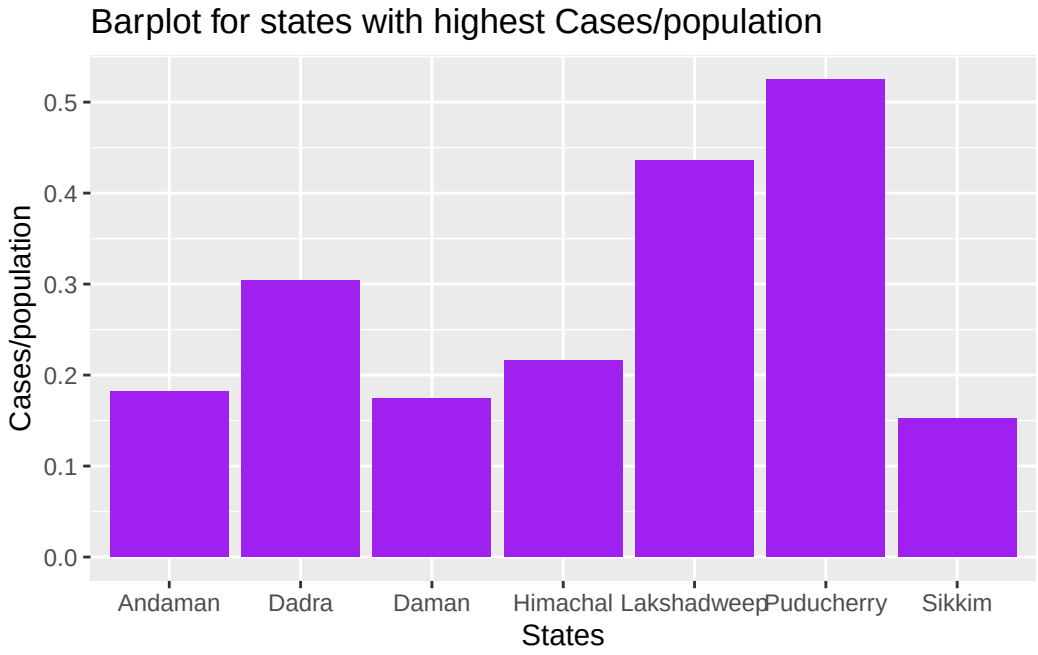

The last thing we took a look at was which states do best in mitigating diseases. There are a lot of possible factors that go into this, including the awareness of the population as well as healthcare facilities in the state. The way we decided to investigate this was to take a look at which the ratio of deaths due to ARIs by the number of cases of ARIs. A fair assumption to

make would be that states which allow lesser deaths in proprtion to the number of cases they see do better in mitigating illnesses.

Upon investigating this we saw that the number of states with the highest cases/population weren't the same as the ones with the highest deaths/cases, probing interesting questions as to why certain states do poorly in dealing with these illnesses.




Assam, Bihar, Chandigarh, Dadra and Nagar Haveli, Daman and Diu , Gujarat, Lakshadweep and Nagaland showed the lowest number of deaths to ARIs (0)


Barplot for states with highest Cases/population
Cases/population
0.0
0.1
0.2
0.3
0.4
0.5
Andaman
Dadra
Daman
Himachal
Lakshadweep
Puducherry
Sikkim
States