**AIR POLLUTION ANALYSIS**

A Mini Project Report Submitted by

Nivedita Suresh Kumar Pooja Mastappa Naik

(4NM17CS116) (4NM17CS125)

Prathika S Shetty Priya Shetty Raksha Shetty

(4NM17CS131) (4NM17CS139) (4NM17CS142)

UNDER THE GUIDANCE OF

**Mrs. Savitha Shetty**

ASST. PROFESSOR GRADE II

Department of Computer Science and Engineering

in partial fulfilment of the requirements for the award of the Degree of

Bachelor of Engineering in

Computer Science & Engineering

From

Visvesvaraya Technological University, Belgaum



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

N.M.A.M. INSTITUTE OF TECHNOLOGY

(An Autonomous Institution under VTU, Belgaum) (AICTE approved, NBA Accredited, ISO

9001:2008 Certified) NITTE -574 110, Udupi District, KARNATAKA.

**April 2020**

(ISO 9001:2015 Certified), Accredited with ‘A’Grade by NAAC

(: 08258 - 281039 –281263, Fax: 08258 –281265

**Department of Computer Science and Engineering**

B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

CERTIFICATE

“**Air Pollution Analysis**”

is a bonafide work carried out by

Nivedita Suresh Kumar Pooja Mastappa Naik

(4NM17CS116) (4NM17CS125)

Prathika S Shetty Priya Shetty Raksha Shetty

(4NM17CS131) (4NM17CS139) (4NM17CS142)

in partial fulfilment of the requirements for the award of

Bachelor of Engineering Degree in Computer Science and Engineering

prescribed by Visvesvaraya Technological University,

Belgaum during the year 2019-2020.

It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report.

The Mini project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Bachelor of Engineering Degree.

Signature of Guide Signature of HOD

# **ACKNOWLEDGEMENT**

We believe that our project will be complete only after we thank the people who have contributed to make this project successful.

First and foremost, our sincere thanks to our beloved principal, **Dr. Niranjan N. Chiplunkar** for giving us an opportunity to carry out our project work at our college and providing us with all the needed facilities.

We express our deep sense of gratitude and indebtedness to our guide **Mrs. Savitha Shetty**, Assistant Professor, Department of Computer Science and Engineering, for her inspiring guidance, constant encouragement, support and suggestions for improvement during the course of our project.

We sincerely thank **Dr. K.R. Udaya Kumar Reddy**, Head of Department of Computer Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology, Nitte.

We also thank all those who have supported us throughout the entire duration of our project.

Finally, we thank the staff members of the Department of Computer Science and Engineering and all our friends for their honest opinions and suggestions throughout the course of our project.

Nivedita Suresh Kumar (4NM17CS116)

Pooja Mastappa Naik (4NM17CS125)

Prathika S Shetty (4NM17CS131)

Priya Shetty (4NM17CS139)

Raksha D Shetty (4NM17CS142)

**ABSTRACT**

Air pollution has been one of the significant problems to deal with in all the nations today. In South Asia, it is ranked as the sixth most dangerous killer. Keeping this in mind, we‘ve analyzed the air quality data of a few Indian States to find some underlying principles or patterns which might give us an insight into how severe the problem is. The approach is purely data -driven and makes use of Big Data Analytics technologies.

Big data is a new driver of the world‘s economic and societal changes. The world’s data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data’s volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the ‘value‘ in the data through Big Data Analytics technologies. Big Data Analytics poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse and complex on a massive scale .

It is interesting to see how data analysis and the day to day instances are coherent and how data analysis can be used to deal with significant problems. Big data has totally changed and revolutionized the way business and organizations work. Leveraging Big Data for environmental protection is still an untapped frontier. In the case of air pollution, once when the key insights on the contributing factors of pollution are identified , the management and prediction becomes far easier.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| 1. INTRODUCTION | 5 |
| 1. OBJECTIVE | 5 |
| 1. TOOLS USED | 6 |
| 1. IMPLEMENTATION | 6 |
| 1. RESULTS | 12 |
| 1. CONCLUSION | 44 |
| 1. REFERENCES | 44 |

**INTRODUCTION**

In this modern world where we come across new technologies every day, we don’t thoroughly examine their effect on the nature around us. Our mini project basically gives a glimpse of how these technologies affect the air. For the analysis, we have picked 5 states, created a dataset based on the factors such as Industry count, Birth-rate, Death-rate, Forest-cover, Population, Vehicle-count and so on and conducted analysis on them.

This Air pollution analysis is a Big Data Analysis project where we have analysed a large dataset not capable of being analysed by a typical database or data analysis software like Excel, Weka etc. After the analysis we have displayed analysed data i.e., graphs in web pages which makes it easy for the user to understand. We hope that this makes some impact on the people about the deteriorating air quality.

Big Data has grasped a lot attention from the market trends, equipment-based performance, and other industry elements. Big data, analytical tools and technologies greatly assist in IT decision making. Even the large organizations find it difficult to deal with the larger datasets in terms of manipulating and managing the Big Data. Big data is particularly a troublesome factor in business analytics since the traditional tools and procedures are not designed to search and analyse massive datasets.

Big Data deals with two classes of data sets, namely structured and unstructured. The records obtained from inventories, orders, and customer information contributes to the structured datasets. The unstructured data set can be obtained from the web, social media, and intelligent devices.

**OBJECTIVE**

The primary objectives of this project is to collect data pertaining to the determinants of the analysis of the air pollution level of five different states in India during the years 2008-2012 and thereafter calculate and generate reports on the same based on the real time data collected. The determinants for the analysis of the air pollution level are:

* Population density of the states (persons/ sq km)
* Percentage of birth and death rate of the people living in a state
* Count of vehicles owned per family
* Count of industries in the state
* Count of people admitted in the hospital due to air borne diseases
* The total area covered by forests in a state (sq km)
* The concentration of particulate matter in the air (NO₂,SO₂,PM)

**TOOLS USED**

* **R** - It is a software programming language and software environment for statistical computing and graphics. It is an integrated suite of software facilities for data manipulation, calculation and graphic display. It is an open source application which includes virtually every data manipulation techniques, statistical model and chart. We plan to use R primarily for pre-processing the data, manipulating it into structured form for the analysis and present end results.
* **RStudio -** RStudio is an integrated development environment (IDE) for R language. It is a code editor and development environment. As soon as you create a new script, the windows within your RStudio session adjust automatically so you can see both your script and the results in your console when you run your syntax.R packages are an ideal way to package and distribute R code and data for re-use by others. RStudio includes a variety of tools that makes developing R packages easier and more productive.
* **MongoDB -** MongoDB is a document-oriented NoSQL database used for high volume data storage. Instead of using tables and rows as in the traditional relational databases, MongoDB makes use of collections and documents. Documents consist of key-value pairs which are the basic unit of data in MongoDB. Collections contain sets of documents and function which is the equivalent of relational database tables.
* **CSV -** A CSV is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension. CSV files can be used with any spreadsheet program, such as Microsoft Excel or Google Spreadsheets.

**IMPLEMENTATION**

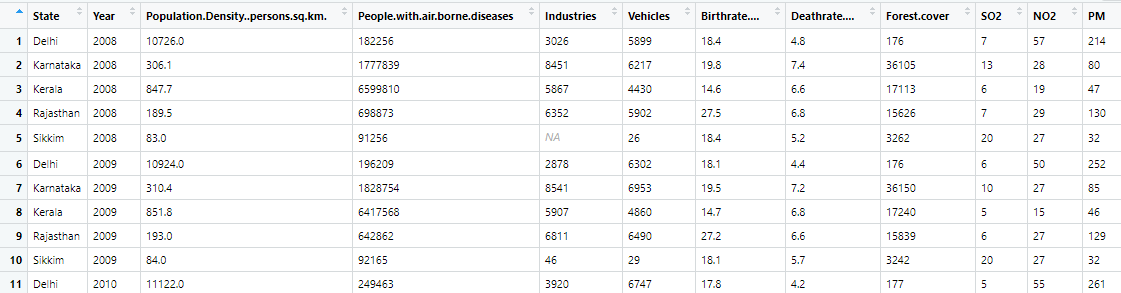
The project is implemented in four stages:

* Collection of data
* Import the data into MongoDB
* Design a data analysis program in R using RStudio to create some visual representation of the data
* Present the results of the analysis using web pages

**Collection Of Data**

The required data of five different Indian states i.e., Delhi, Karnataka, Kerala, Rajasthan and Sikkim over the period of 2008 – 2012 is collected from different sources.

Here is the snapshot of the first few rows of the data collected.



The dataset contains:

* Two categorical variables:
* **State**
* **Year**
* Ten numerical variables:
* **Population.Density..persons.sq.km** – Population per unit area.
* **People.with.air.borne.diseases** – Count of people admitted in hospital due to air

borne diseases.

* **Industries** – Count of Industries.
* **Vehicles** – Count of vehicles in thousands.
* **Birth-rate** - Live births per thousand populations per year.
* **Death rate** - Number of deaths per thousand population per year
* **Forest.cover** – Land area that is covered by forests or the forest canopy or open

woodland in square kilometres.

* **SO2** – Concentration of Sulphur dioxide
* **N02** – Concentration of Nitrogen dioxide
* **PM** – Particulate Matter Concentrations.

The complete dataset has 25 rows and 12 columns and is stored in a csv file named pollution.csv.

**Import The Data Into Mongodb**

MongoDB provides the mongoimport utility that can be used to import JSON, CSV or TSV files into a MongoDB database.

The pollution.csv file is imported into the MongoDB database using the command shown in the figure below.

****

**--db**: Specifies the database to use. We used a database called Pollutionanalysis.

**--collection**: Specifies the collection name. We used a collection called Pollution.

**--type**: The input to import.

**--headerline**: Specifies that the first row in the imported file should be the field names.

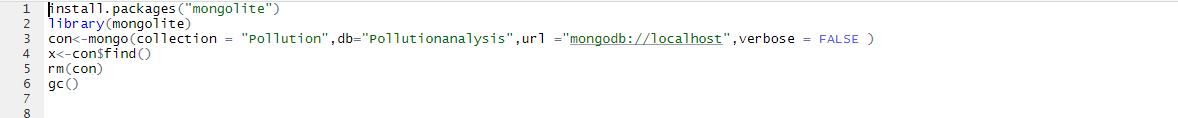
**Design A Data Analysis Program In R Using Rstudio To Create Some Visual Representation Of The Data**

To use MongoDB with R, we can use the mongolite package which is a fast and simple MongoDB client for R.

Package ‘**mongolite**’

**Description:** High performance MongoDB client based on ‘mongo-c-driver’ and ‘jsonlite’. Includes support for aggregation, indexing, map-reduce, streaming, encryption, enterprise authentication and GridFS.

The code in the following figure establishes a connection to MongoDB collection, retrieves the required data and stores it in an object called x.

****

The function **mongo()** supports the following arguments:

* **Collection**: name of the collection to connect to.
* **db**: name of the database to connect to.
* **url**: address of the MongoDB server in standard URI Format.
* **Verbose**: if TRUE, emits some extra output.

The function returns a mongo connection object that acts as a pointer to a collection on the server.

The **find** method automatically simplifies the collection into a data frame. In the above code, it stores the data frame into an object called x.

**rm(con)** followed by **gc()** automatically disables the connection to MongoDB server when all the objects using the connection are removed.

The code in the following figure uses arrange function of dplyr package to reorder the rows in an appropriate order.



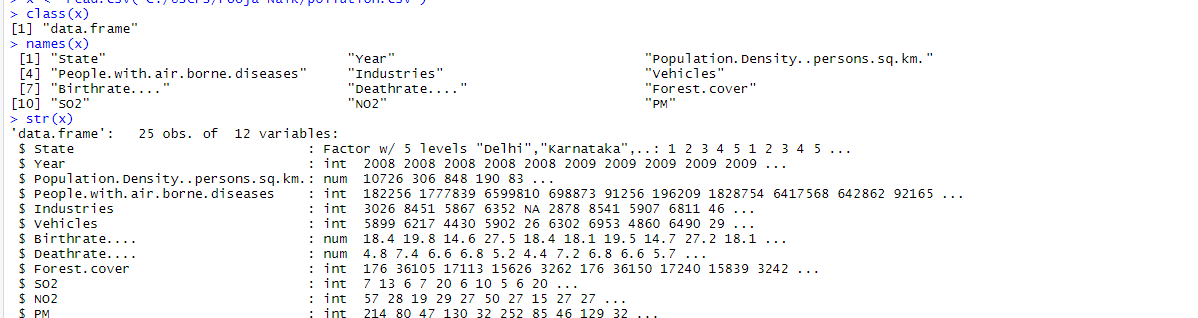
R Studio provides several functionalities to familiarize with the data set.

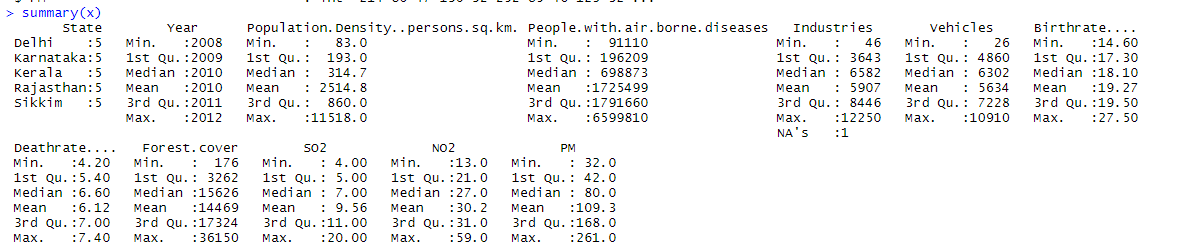
**class()** – Helps in recognizing the class of the data set.

**names()** – This function helps in checking out all the variables in the data set.

**str()** – This helps in understanding the structure of the data set, data type of each attribute and number of rows and columns present in the data.

**Summary()** – One of the most important functions that help in summarizing each attribute in the dataset.

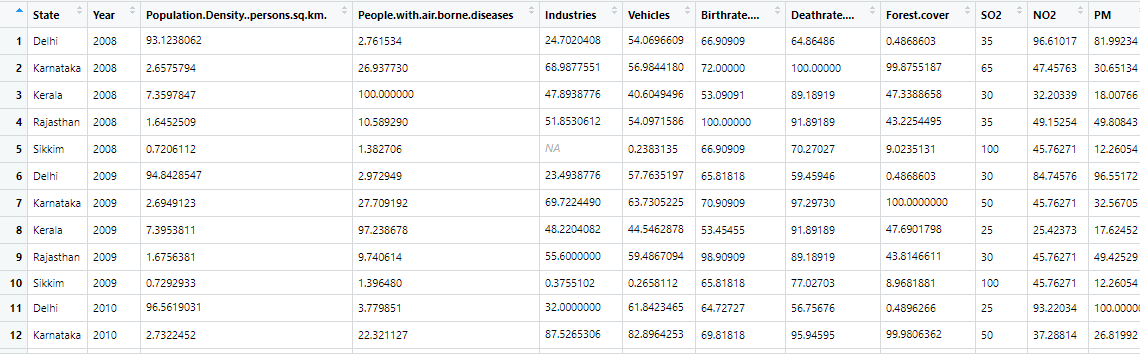




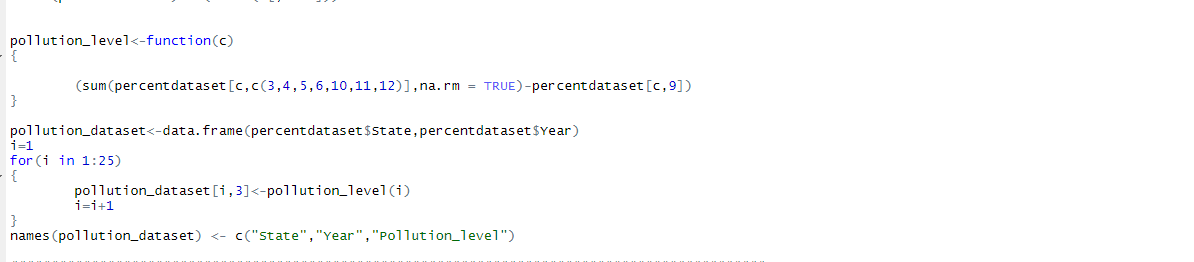
For a better analysis, we have defined a function called percentdata that converts all the values of the data set into the same range. The new values are stored in a data frame called percentdataset.



Here is the snapshot of the first few rows of percentdataset.



A function named pollution\_level is defined to analyze the air pollution level in each state during a particular year using the factors that directly or indirectly play a key role in the pollution level of a place.



A new data frame named pollution\_dataset is created to store the generated pollution level value of each state during a particular year.

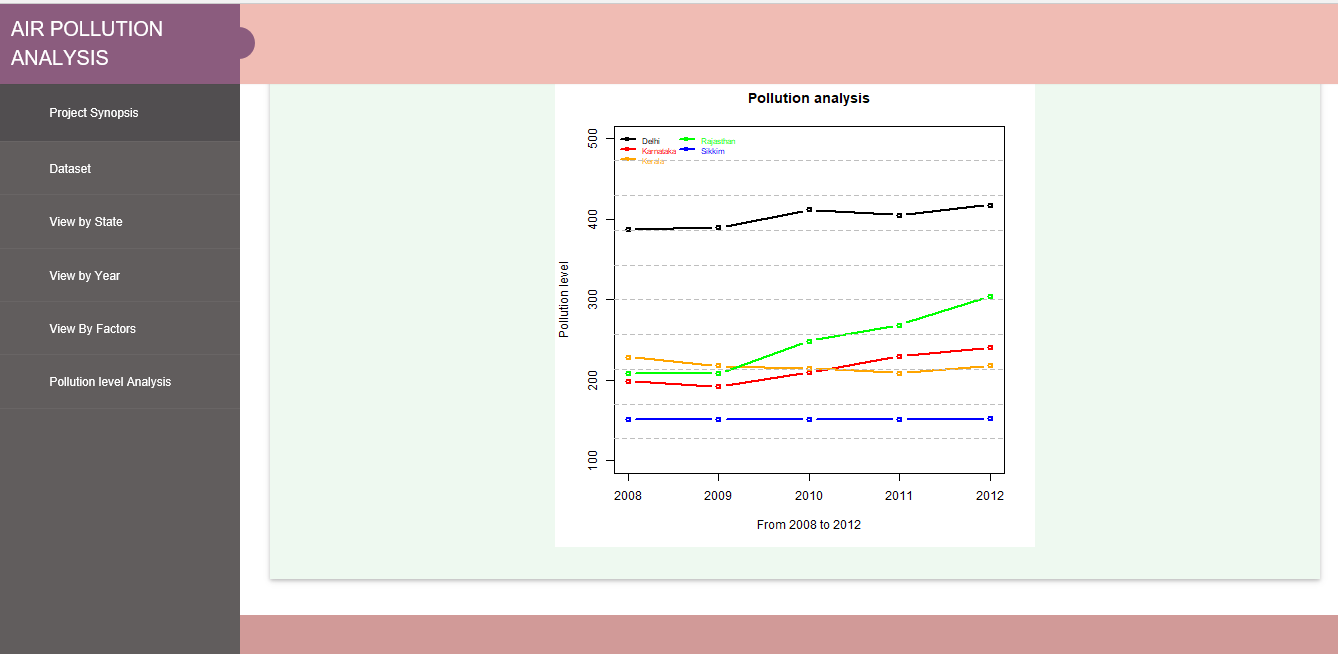
When conducting data analysis, plotting is critically important. The core plotting and graphics engine in R is encapsulated in the following packages:

* **Graphics**: Contains plotting functions for the “base” graphing systems, including plot, hist, boxplot and many others.
* **grDevices**: Contains all the code implementing the various graphics devices, including X11, PDF, PostScript, PNG etc.

**Present The Results Of The Analysis Using Web Pages**

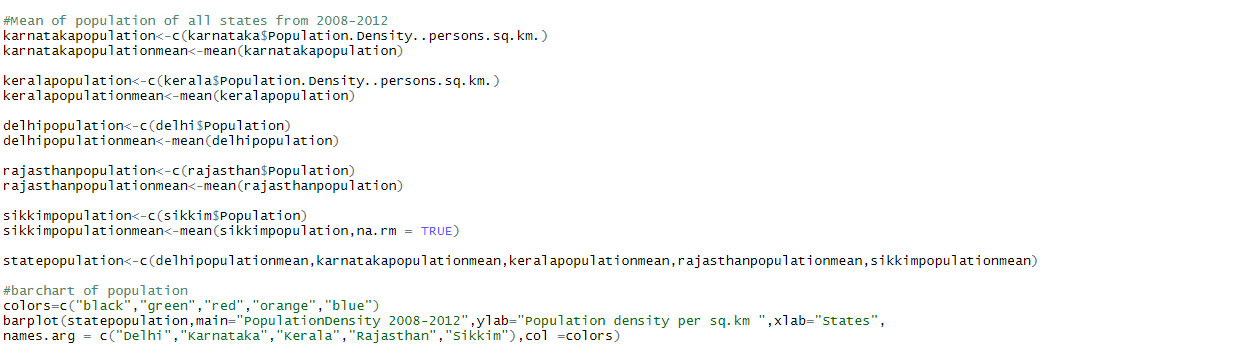
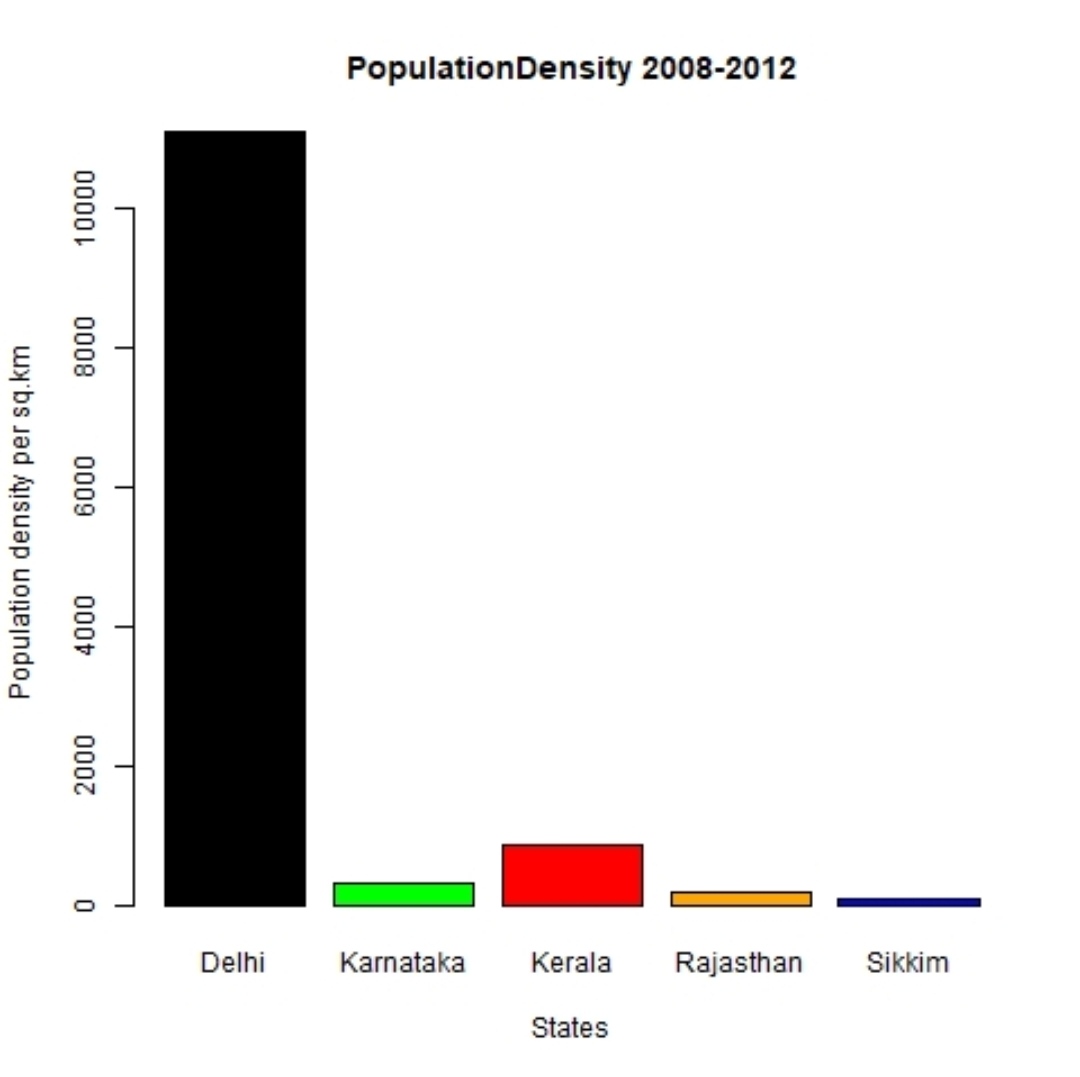
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data.

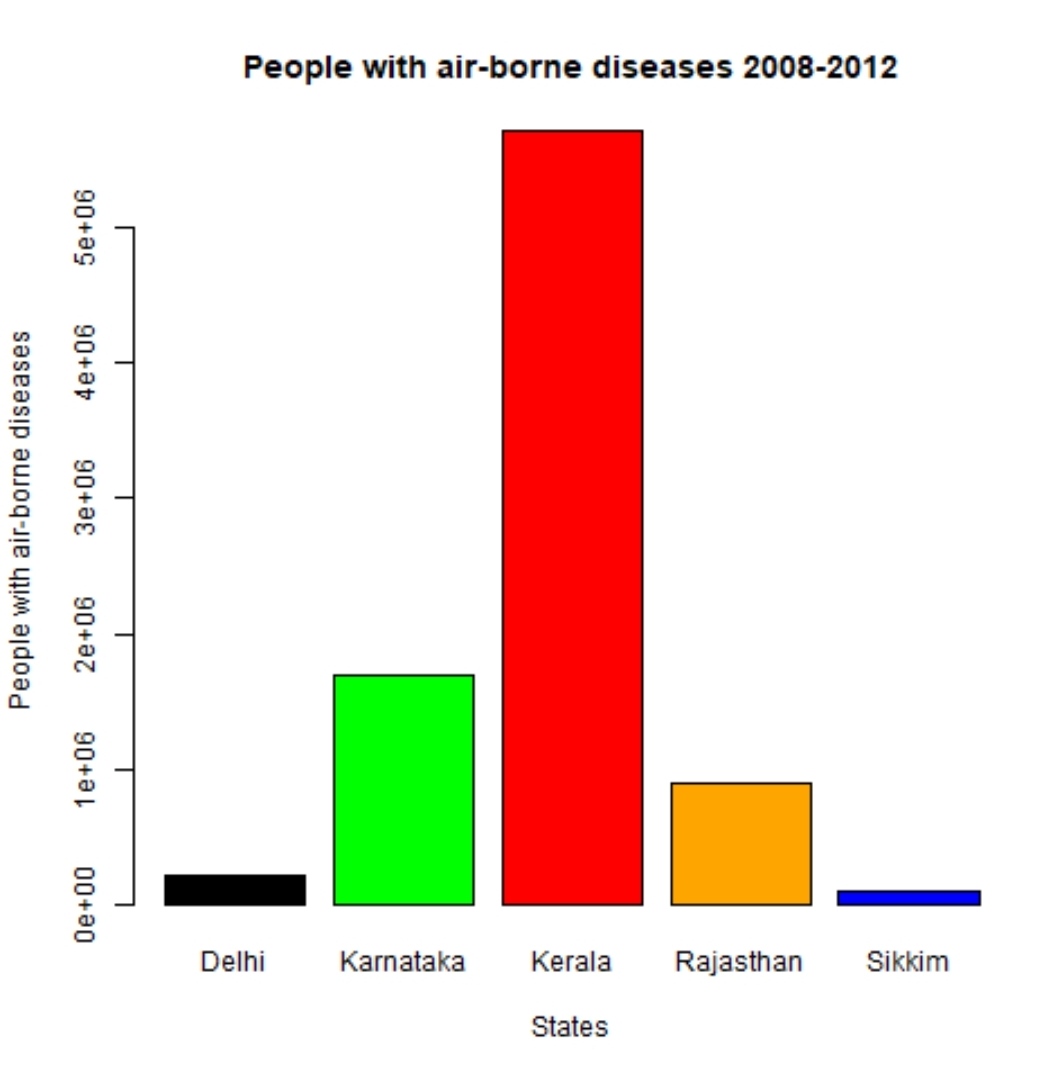
We’ve designed web pages to present the results and conclusions of our analysis in a better way. The web pages contain all the visual elements along with a brief conclusion of what they represent which makes the whole process of analysis more feasible.

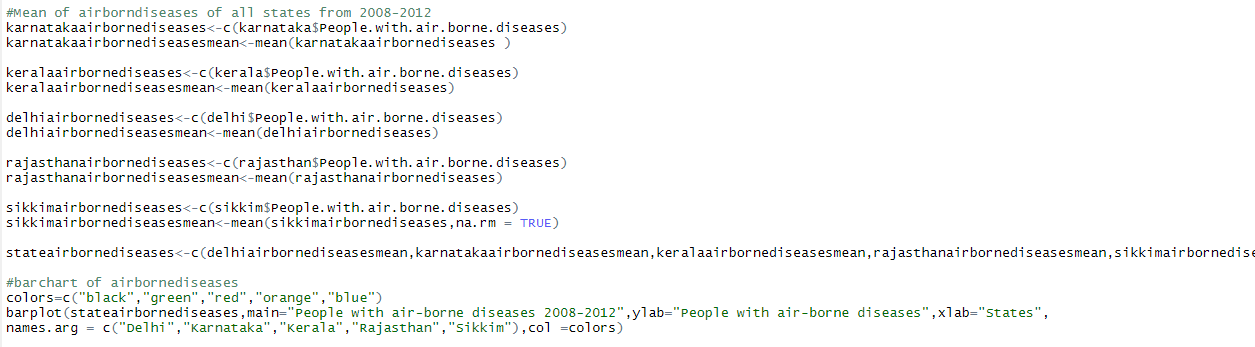


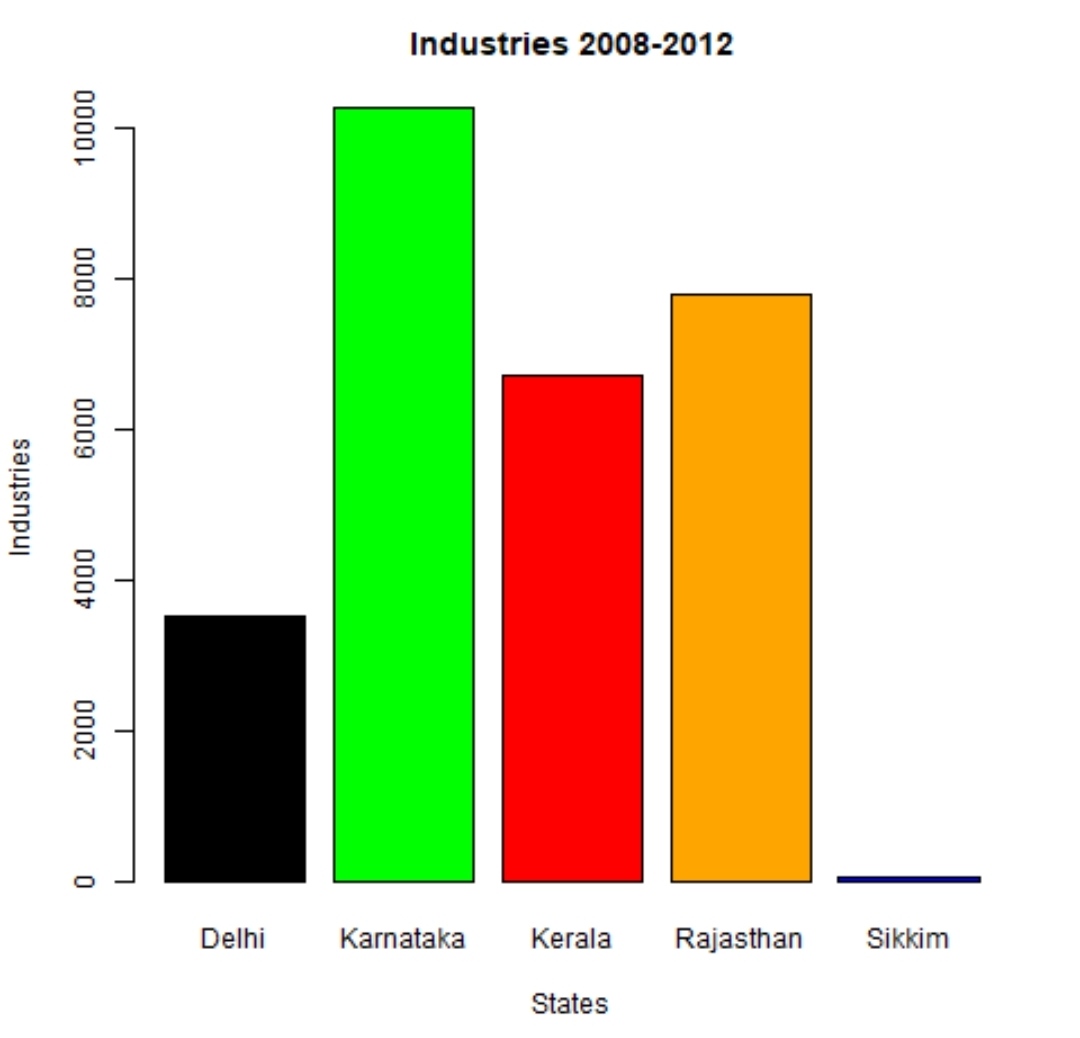
**RESULTS**

The following bar graphs show the variation of each factor in all five states during the period 2008 to 2012.

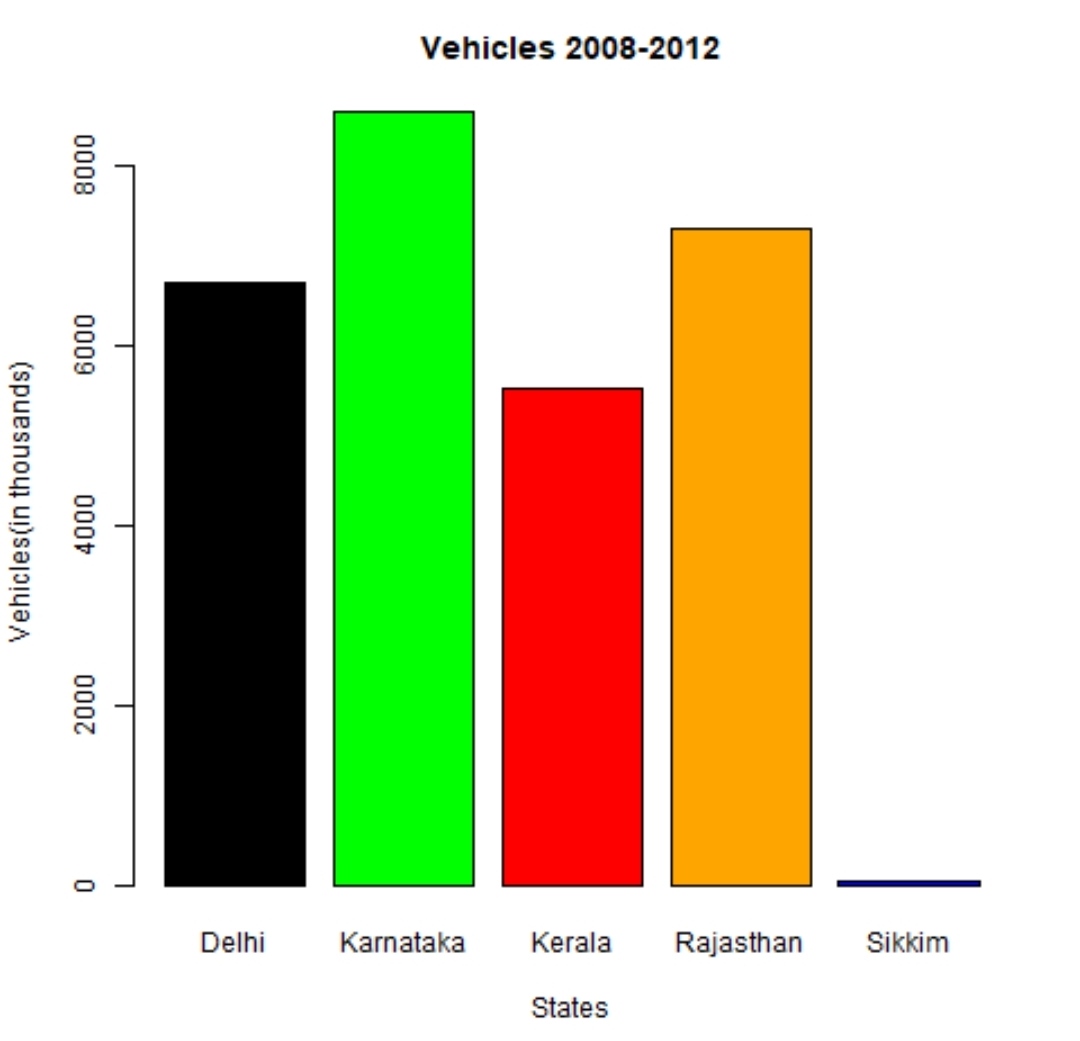




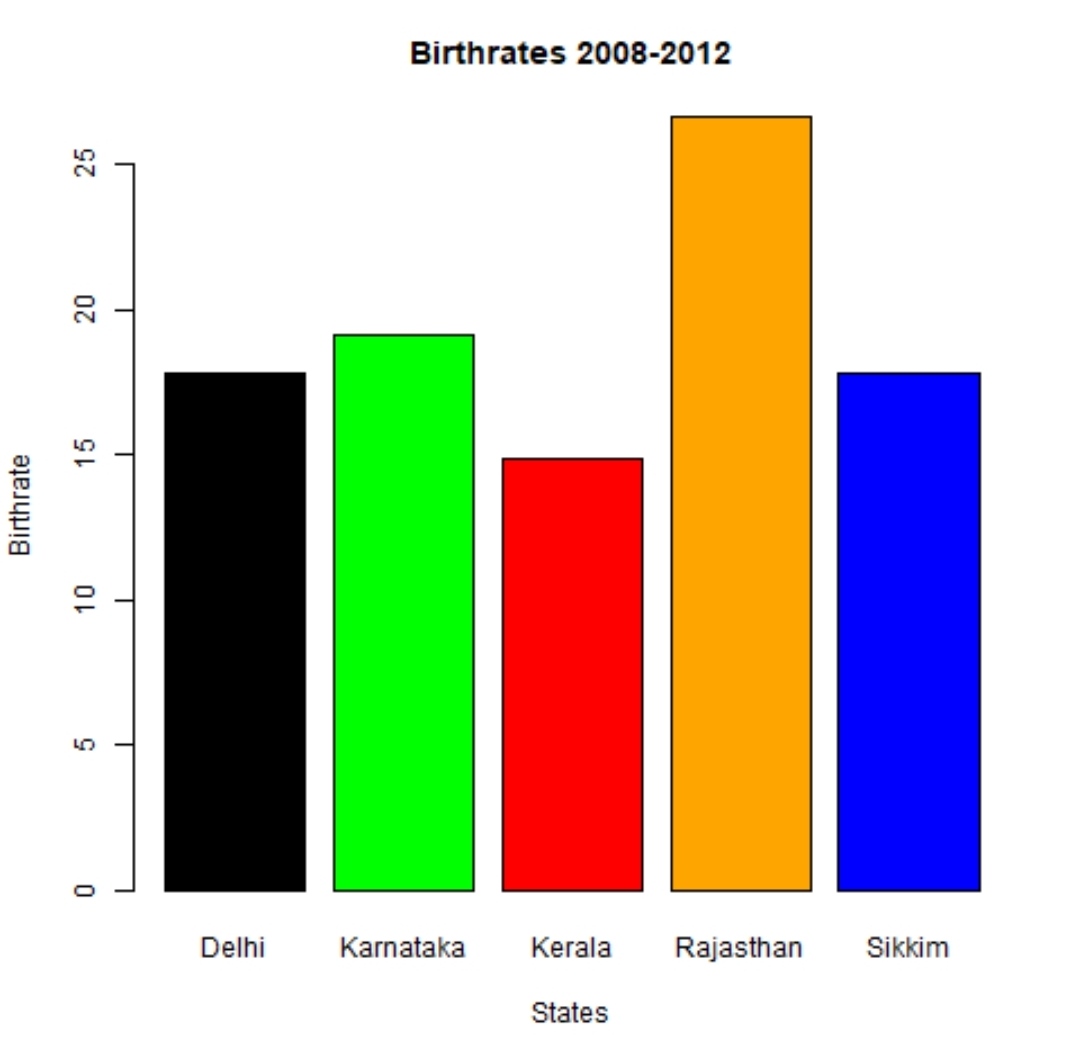


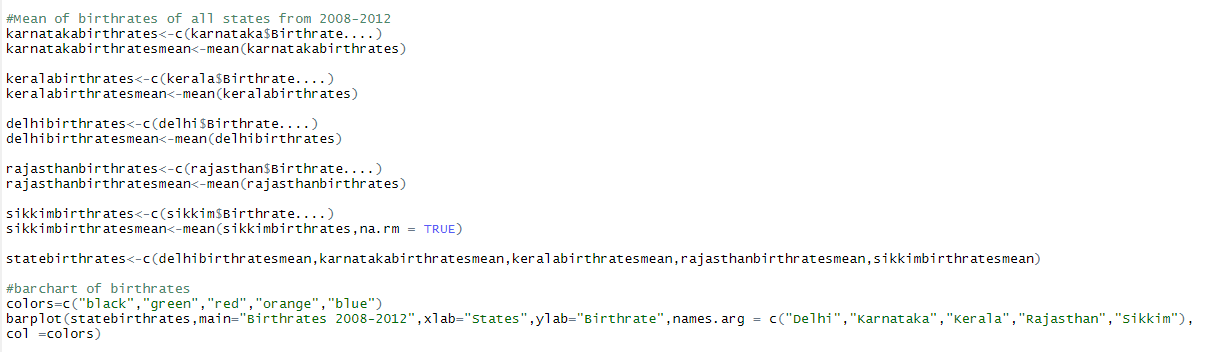


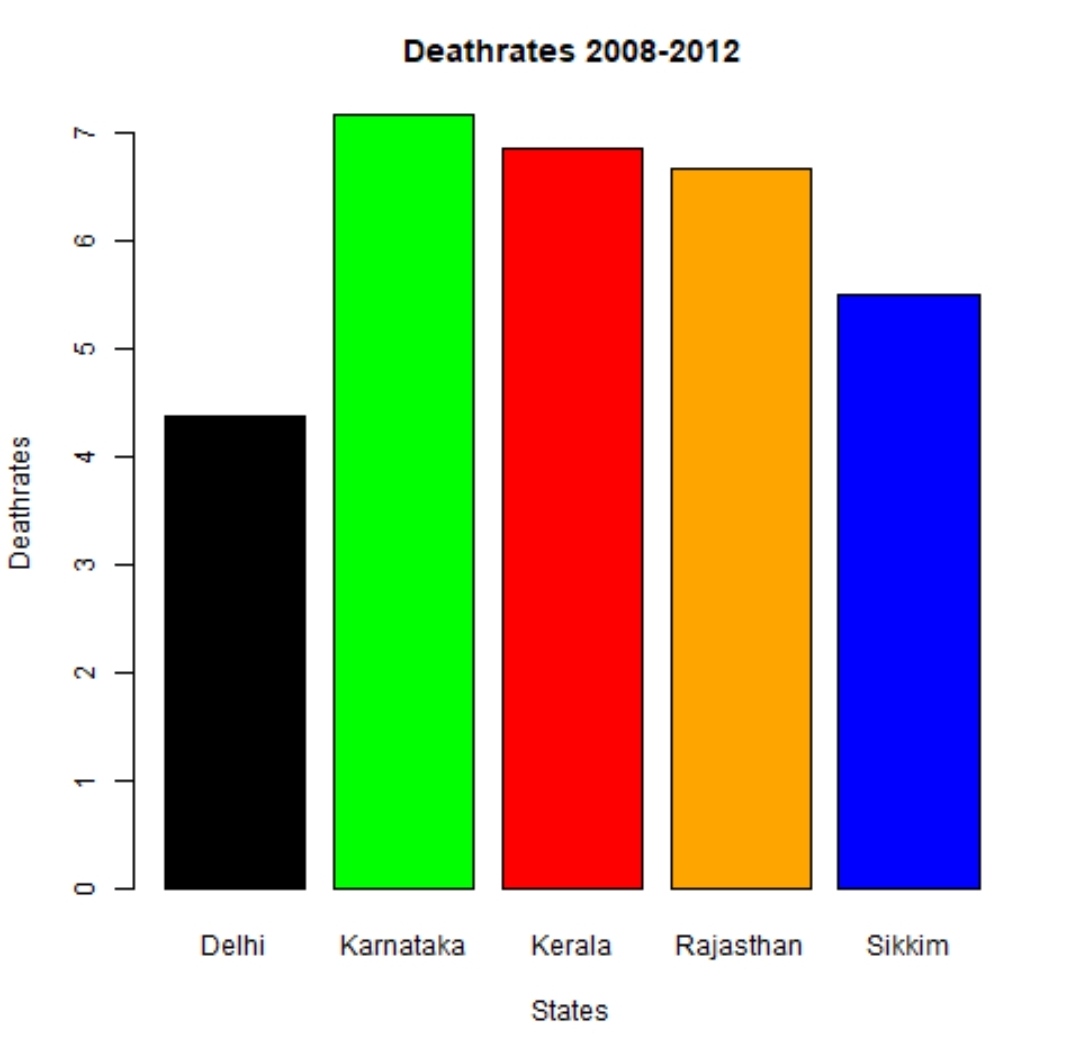


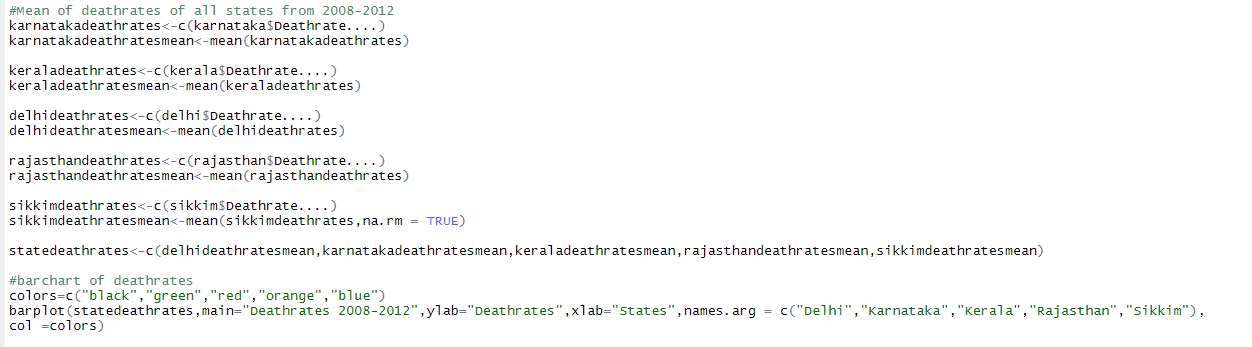


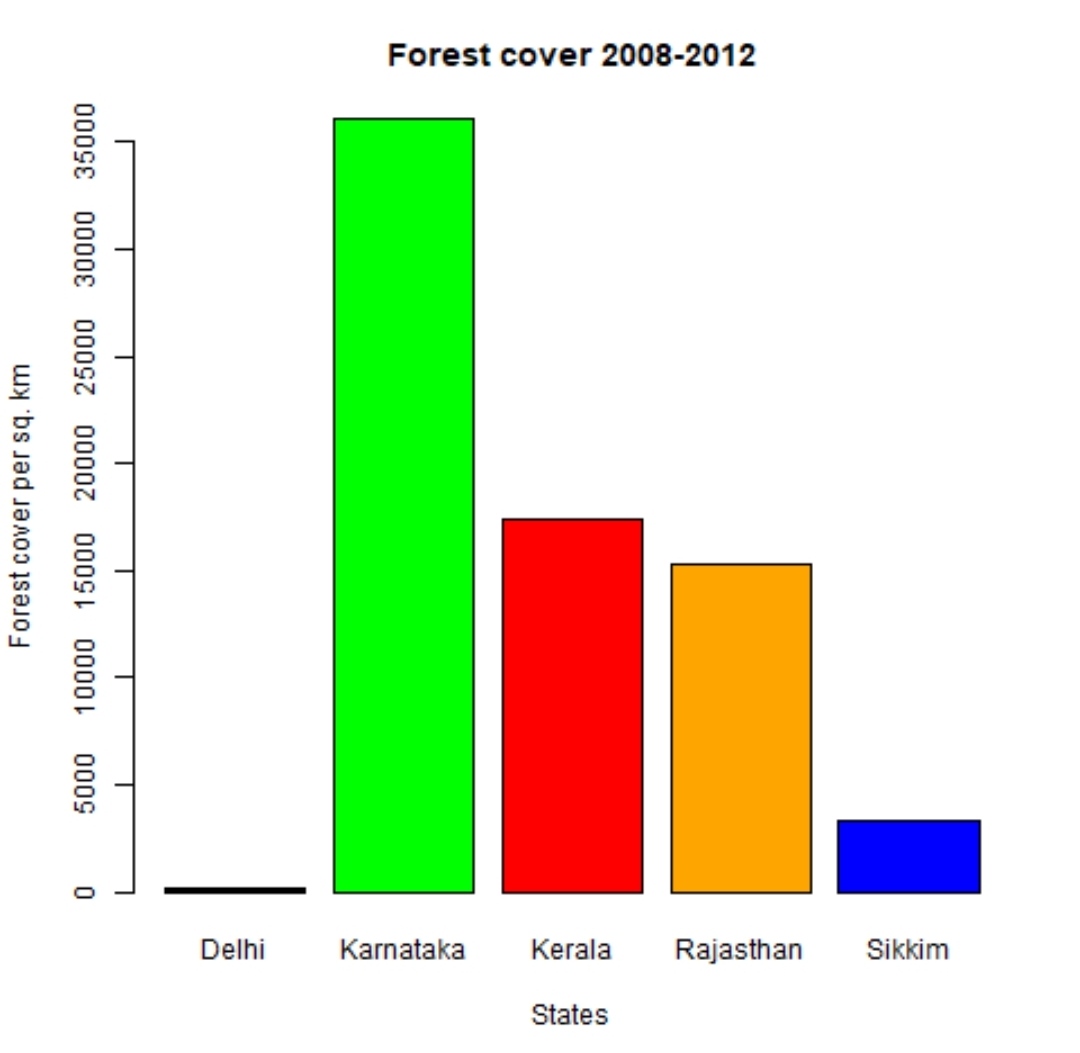




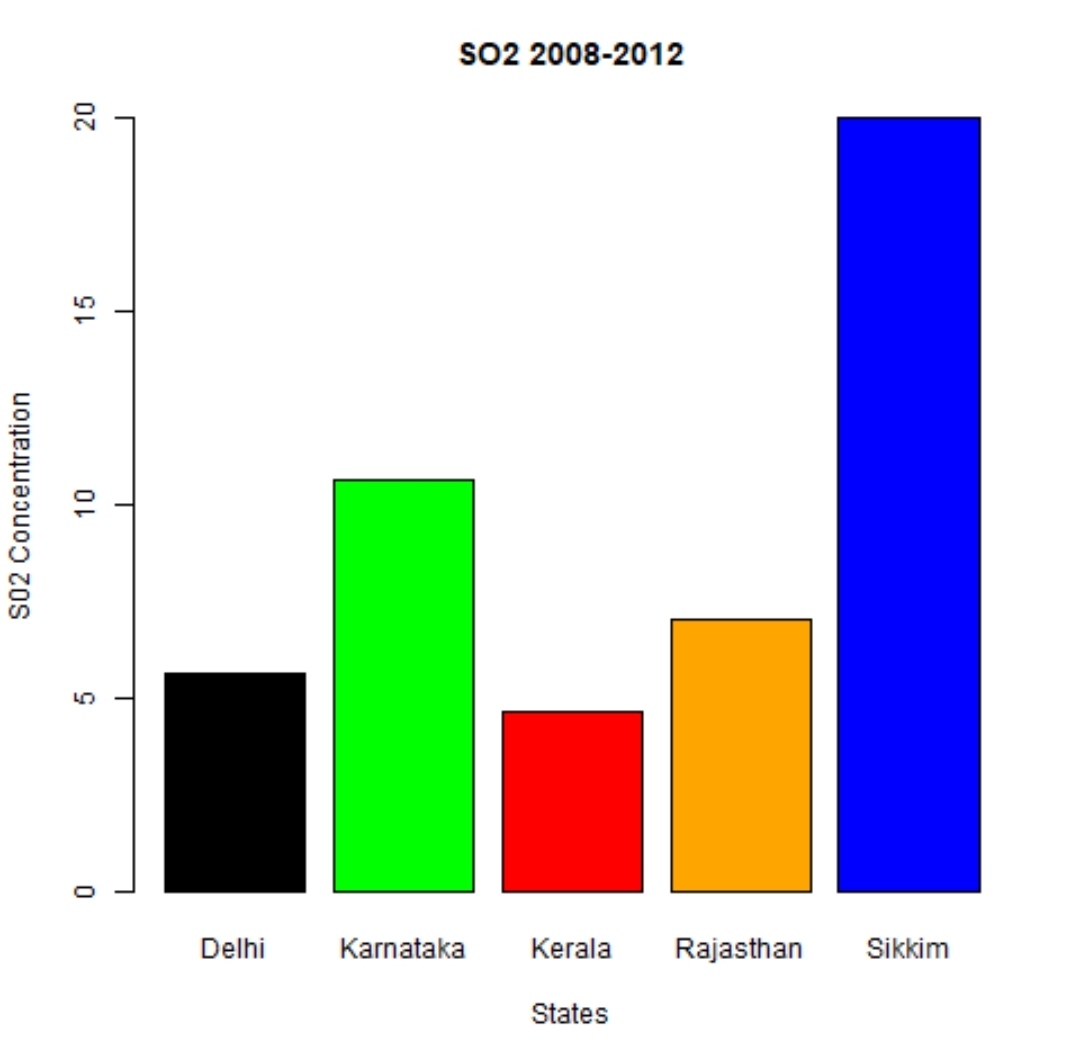


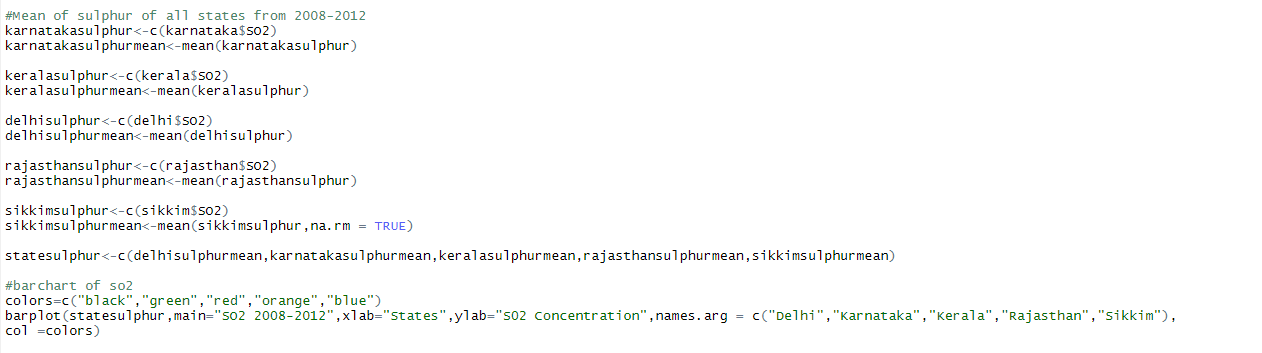


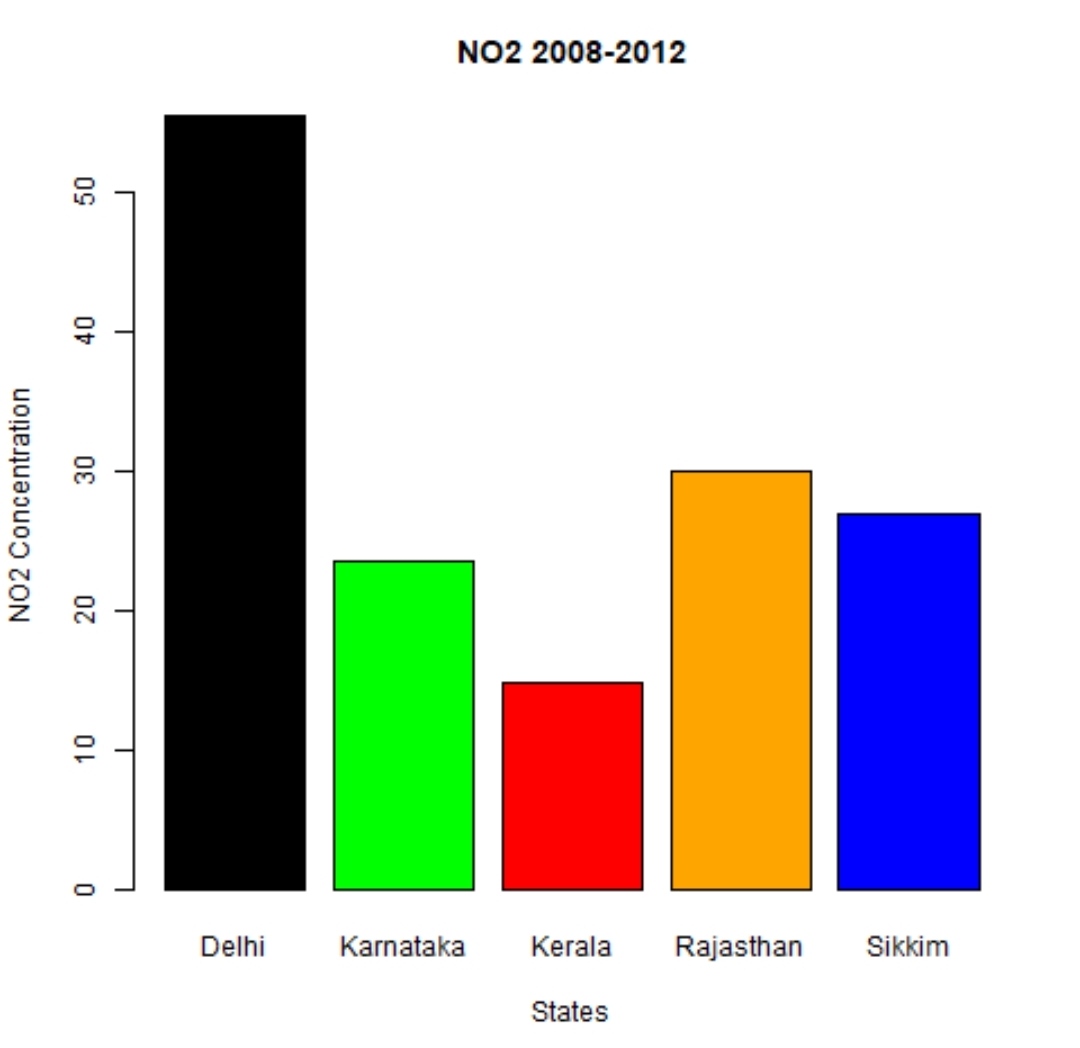


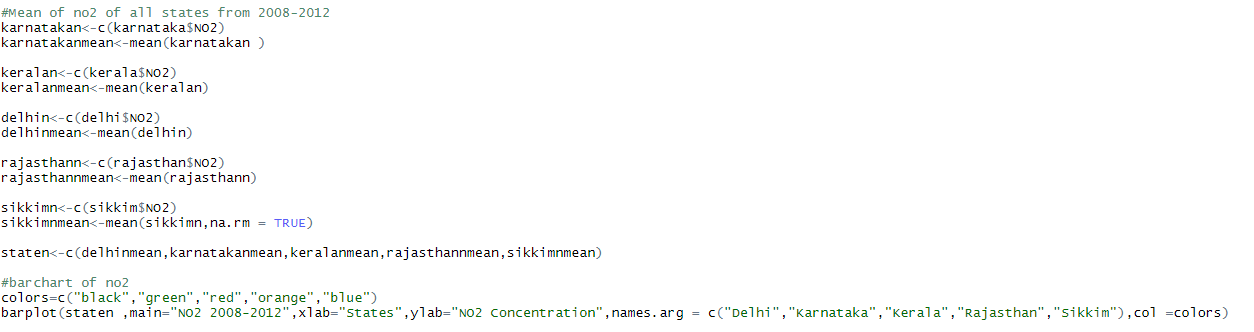








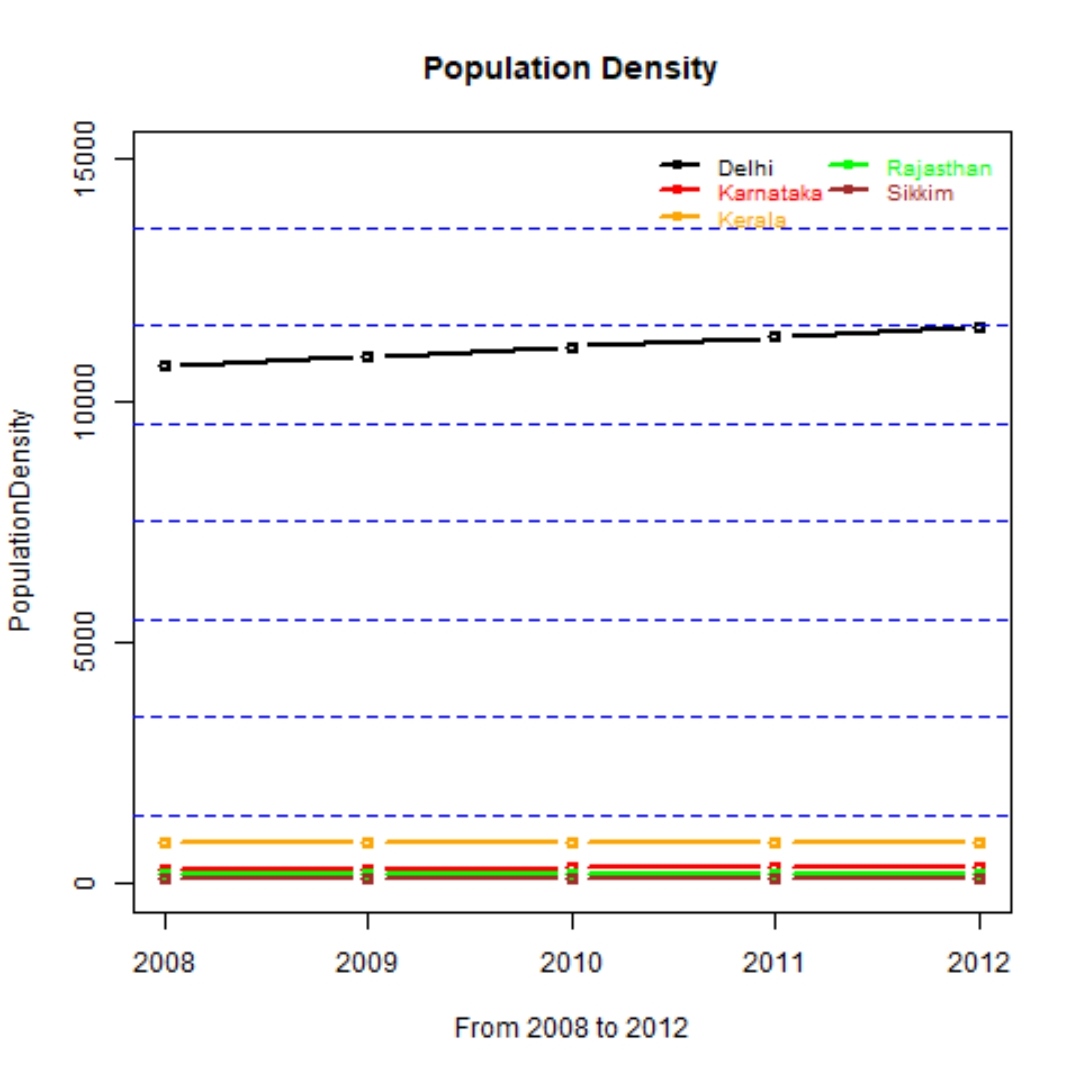




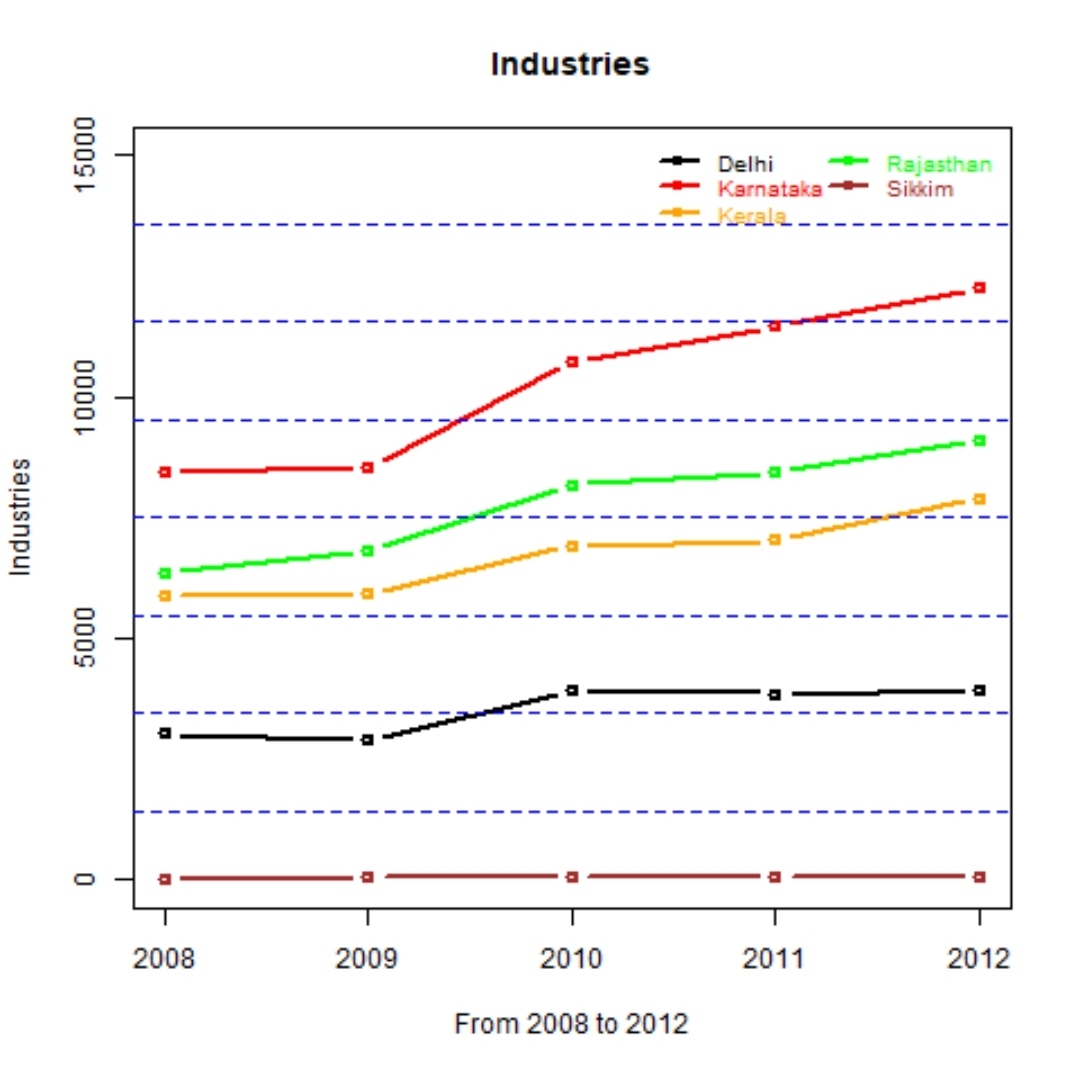


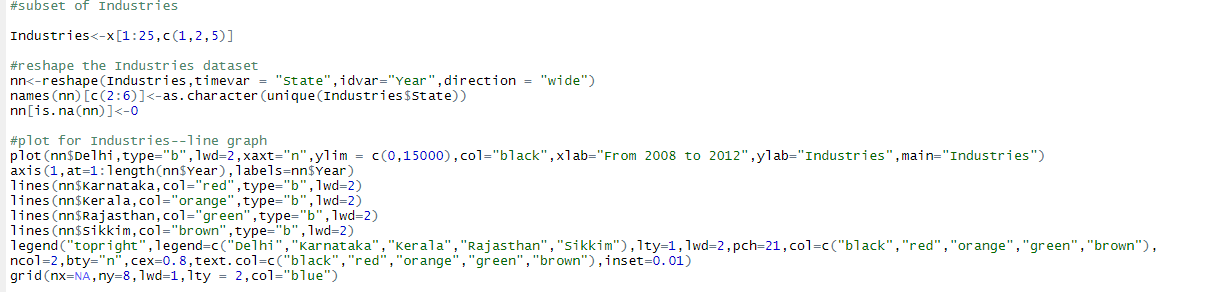


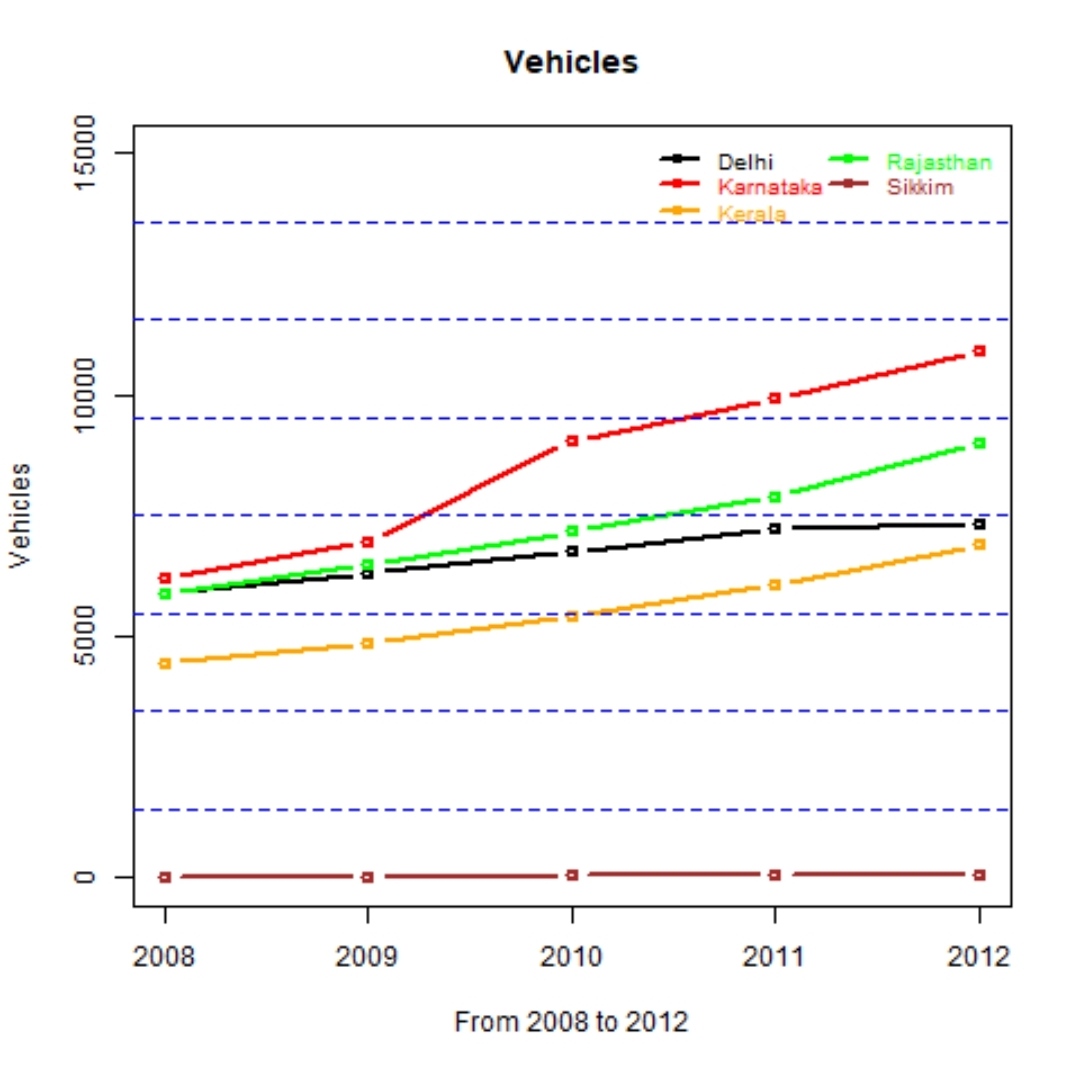
The following line graphs show the variation of each factor in all five states during the period 2008 to 2012.

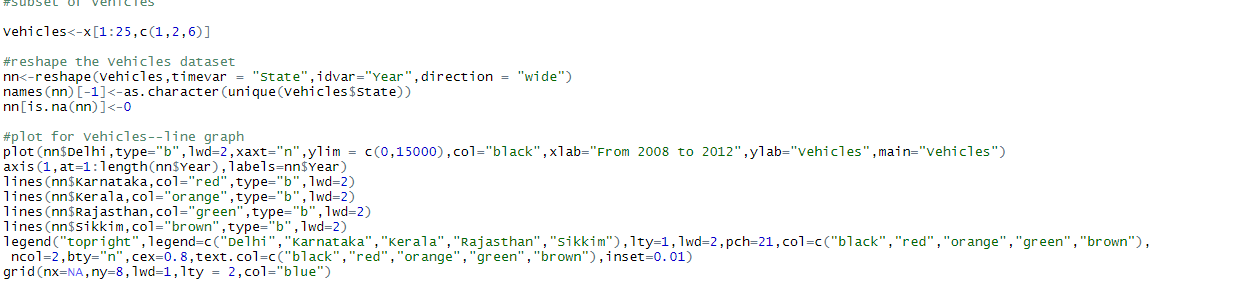


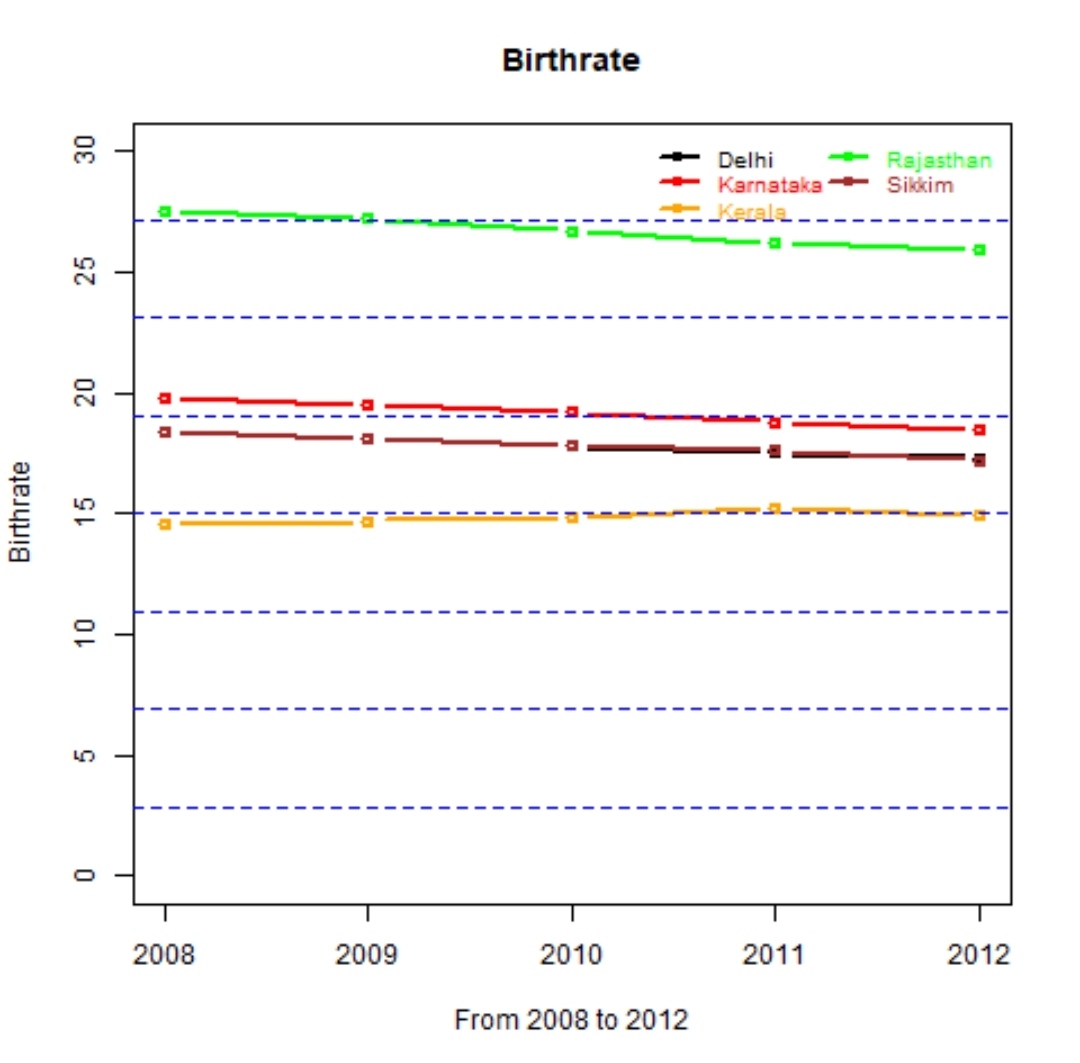


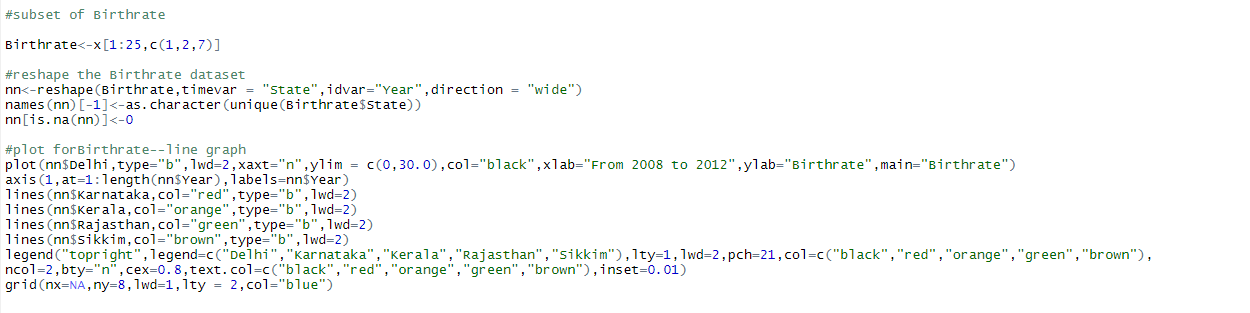


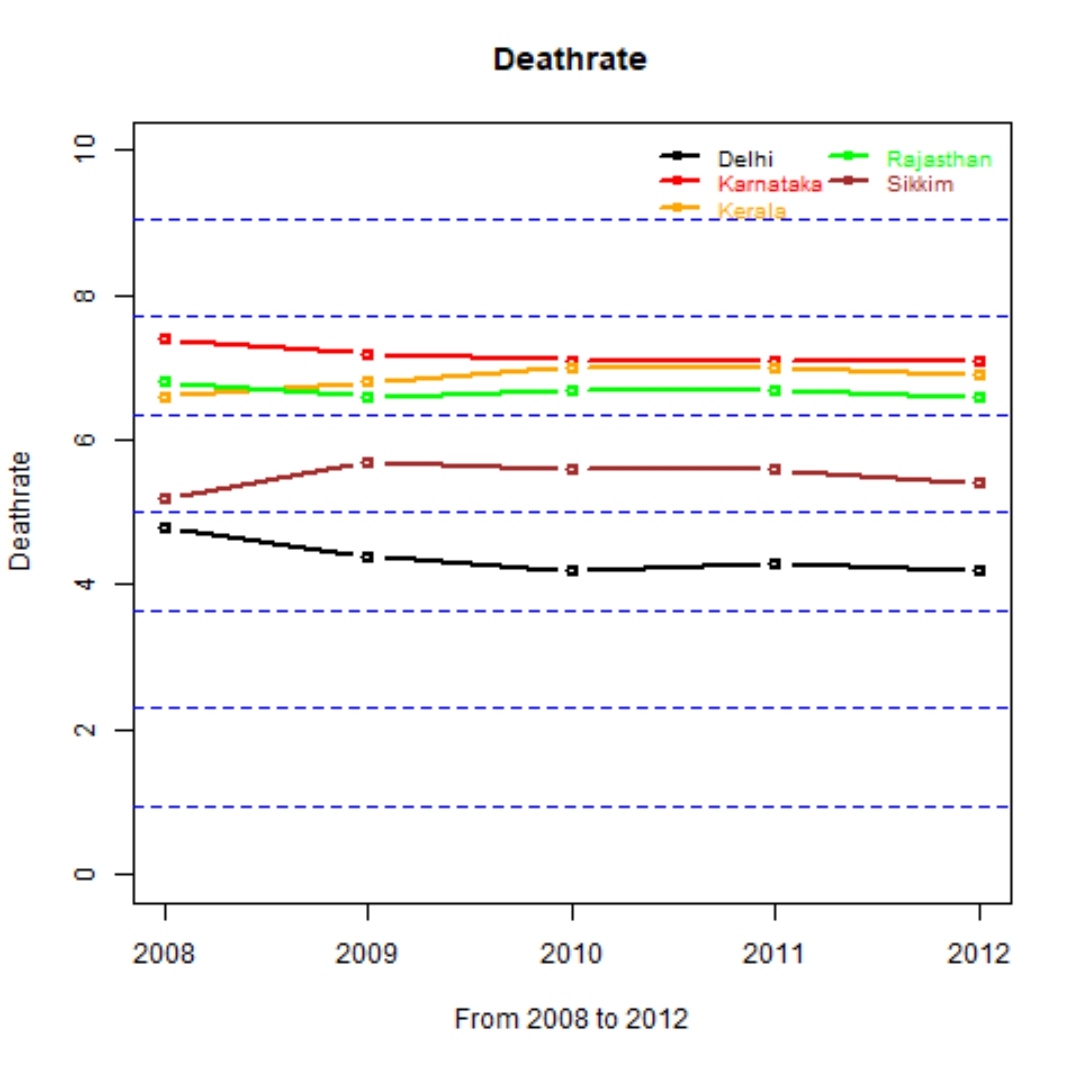


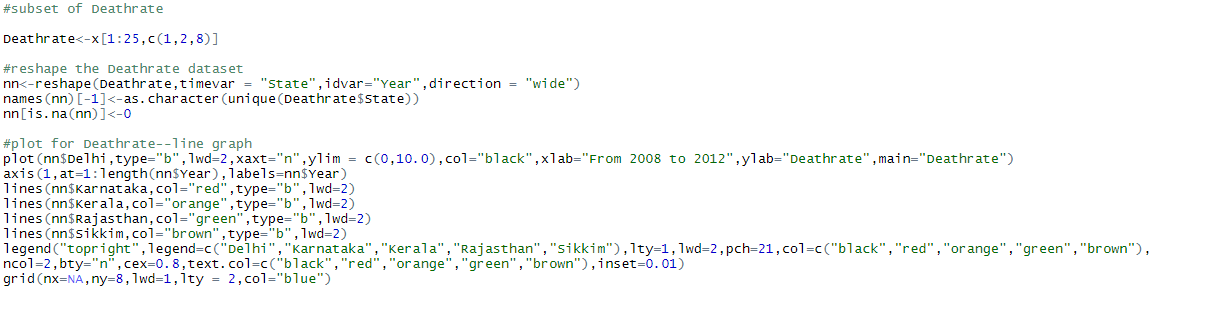


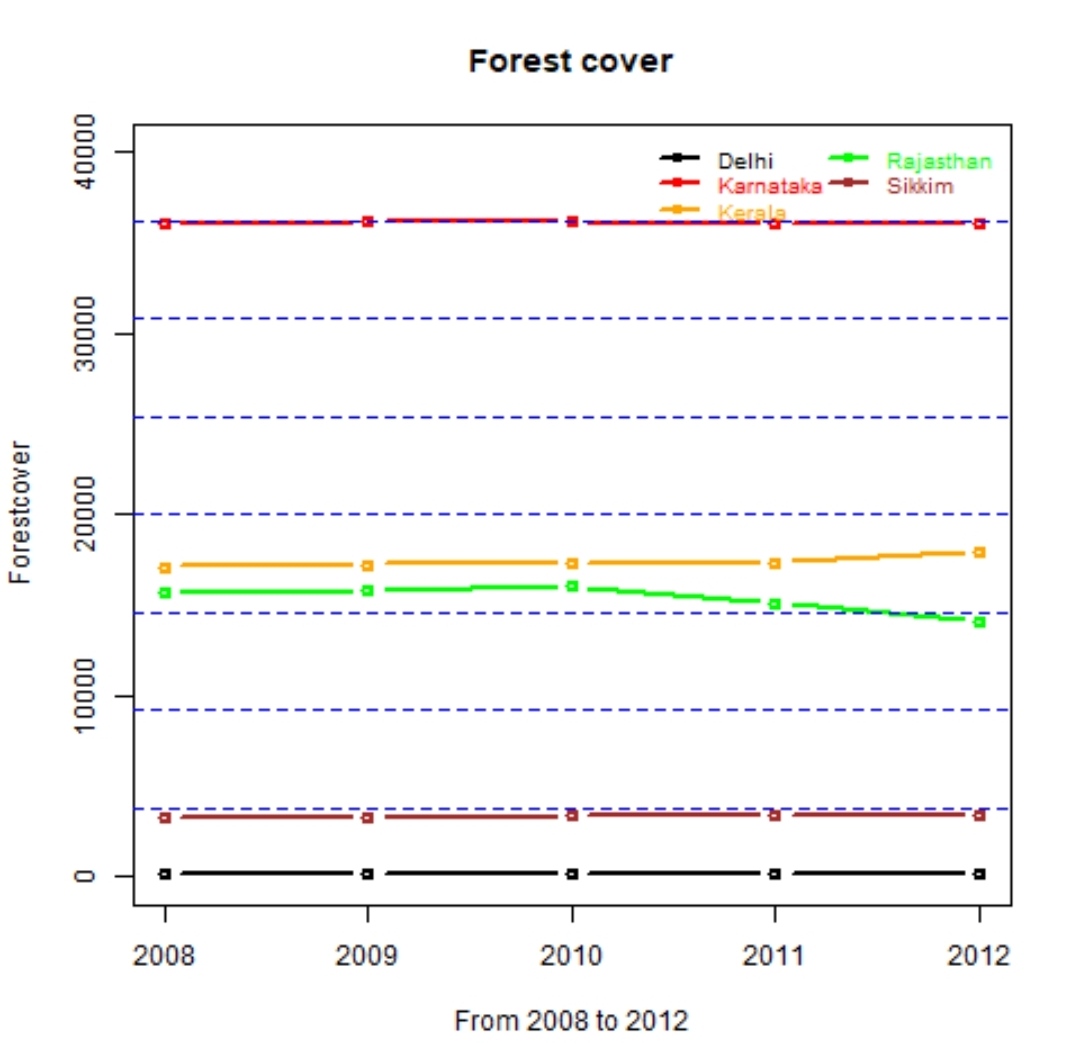






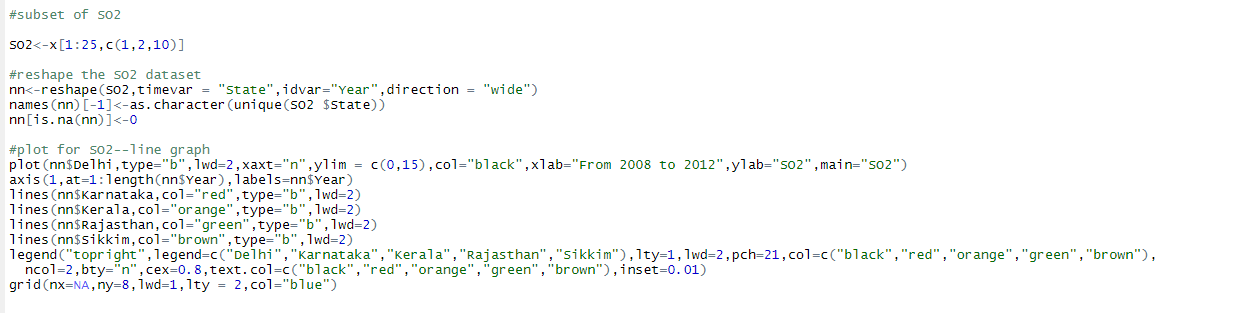


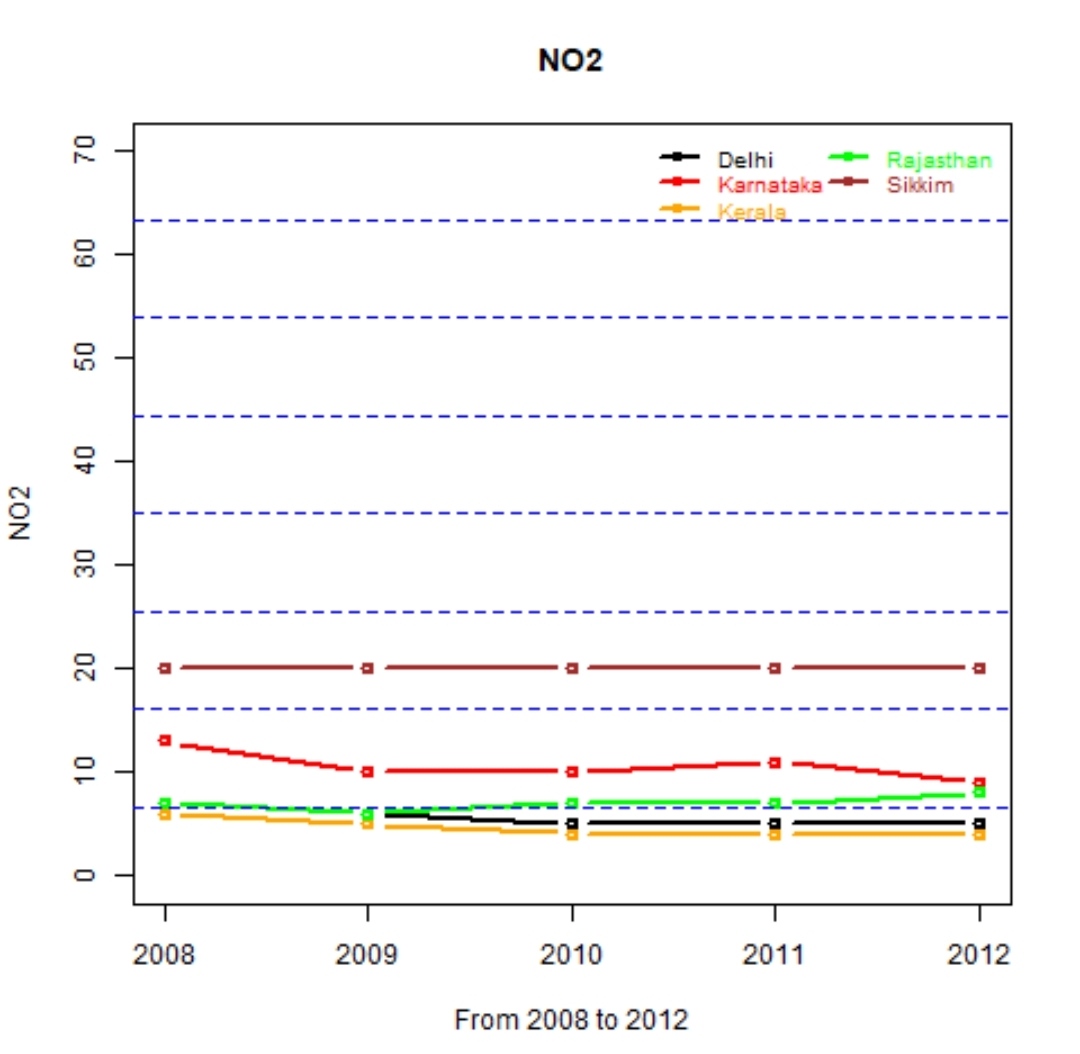


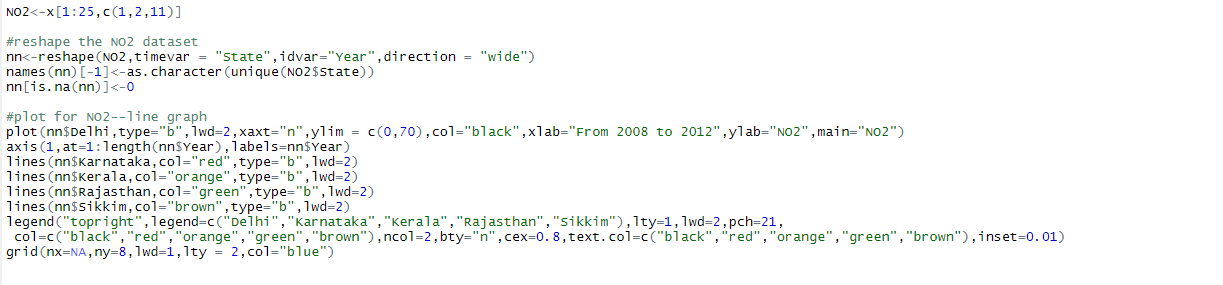


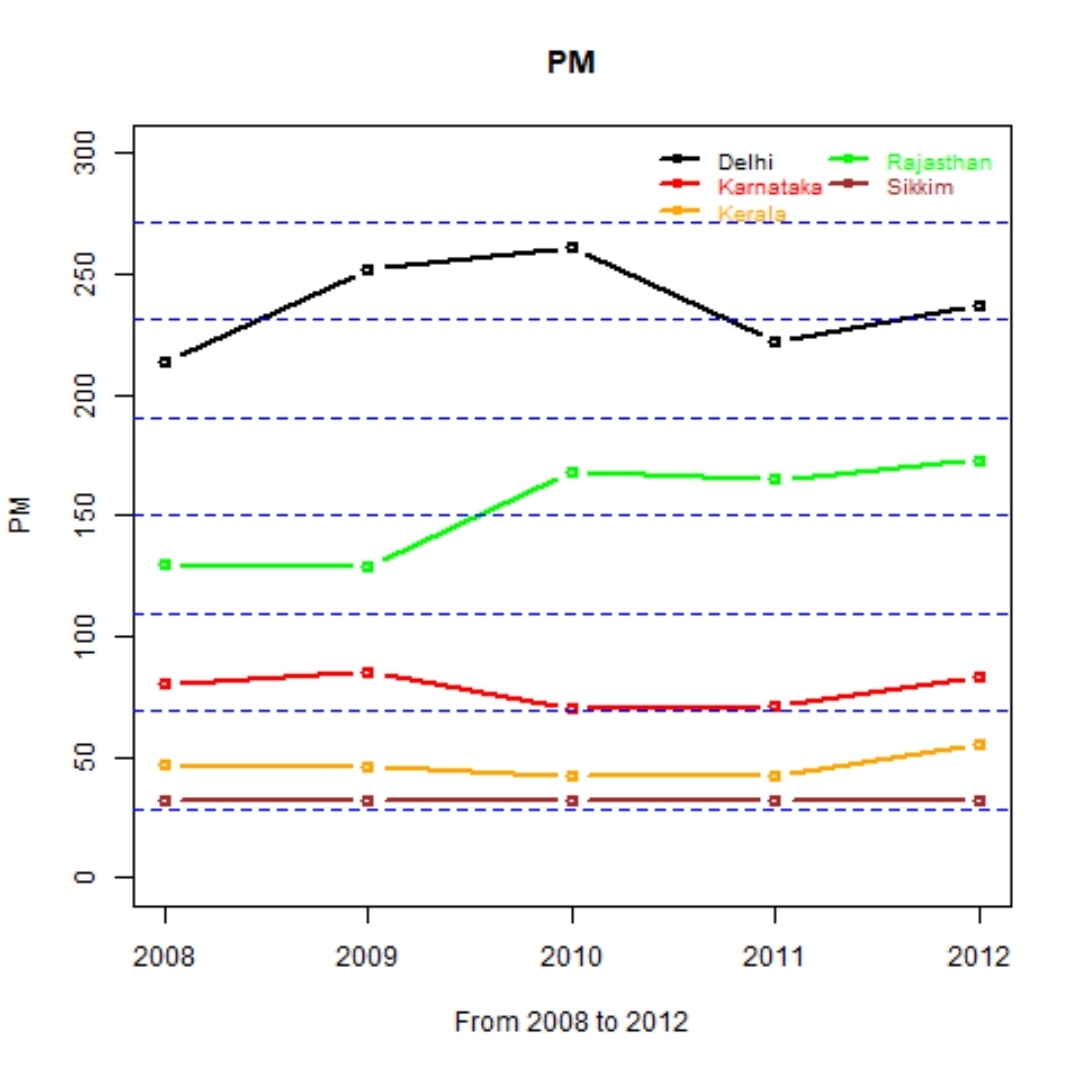


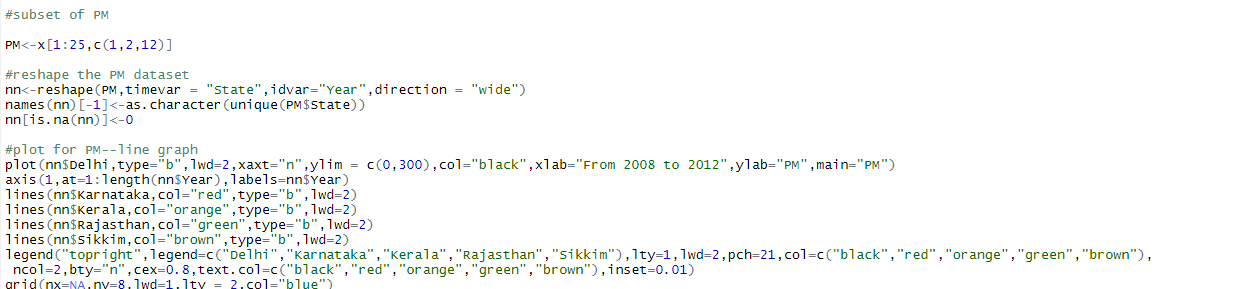




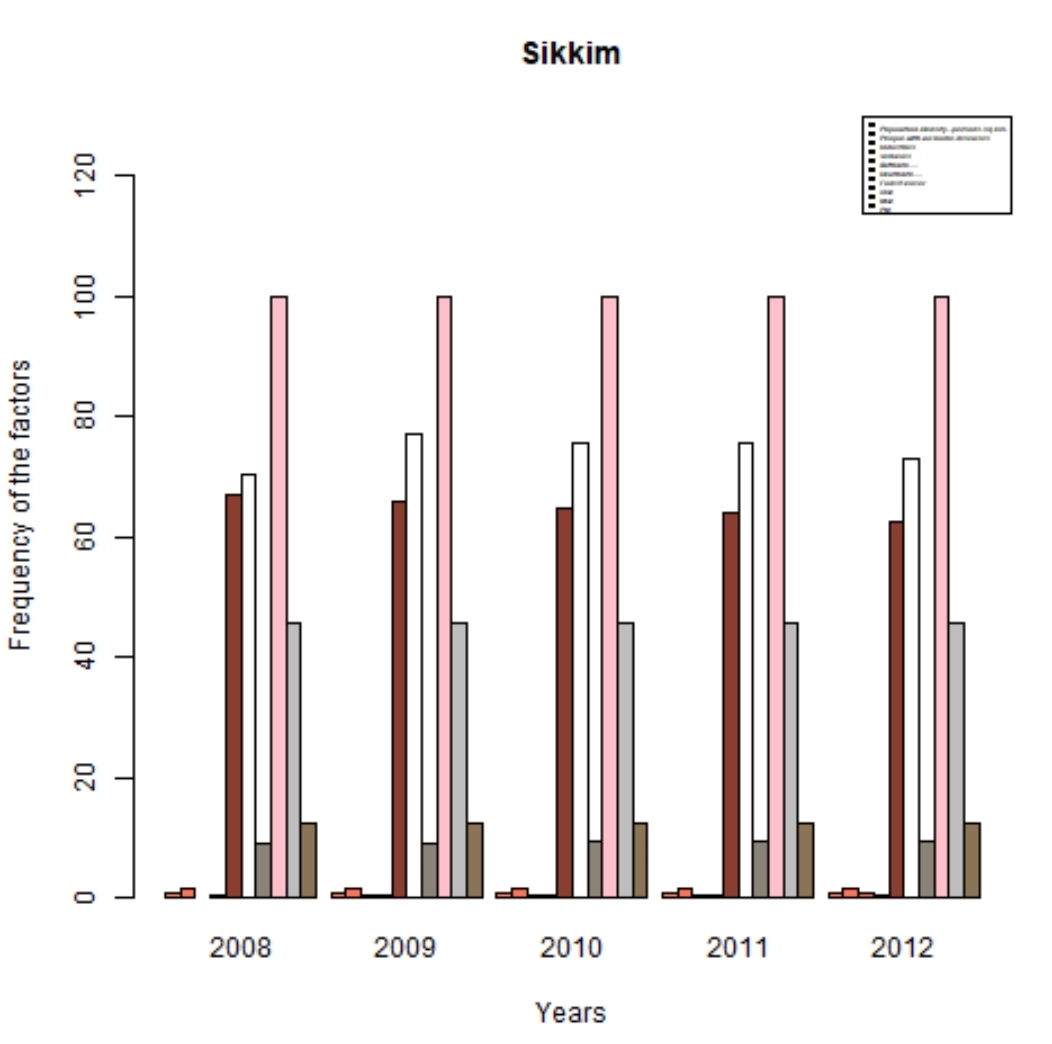


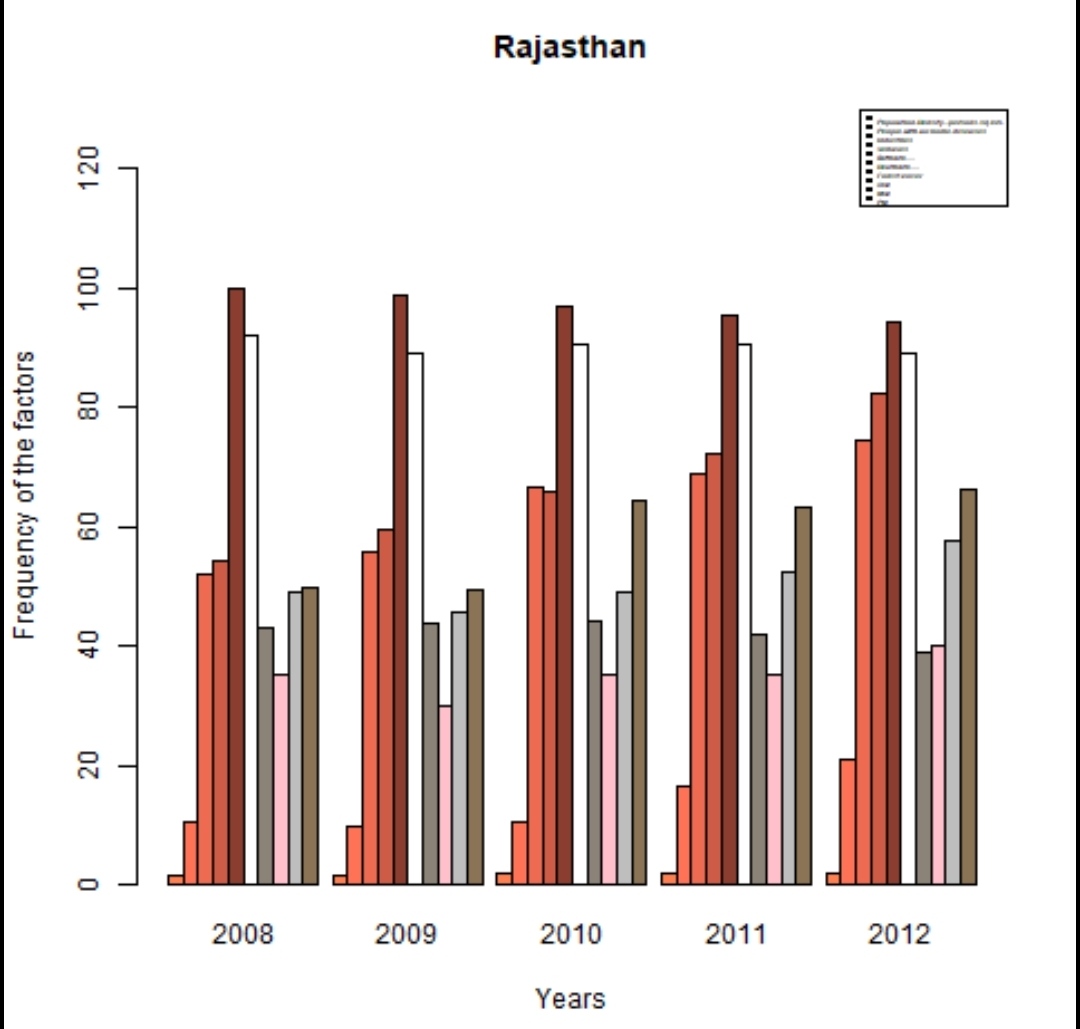


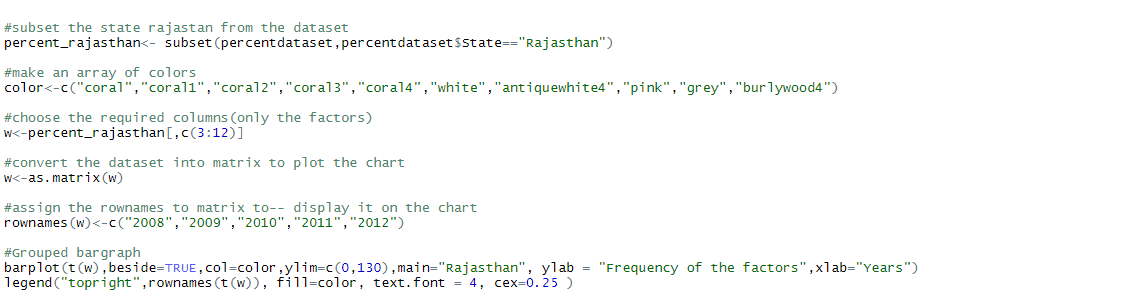


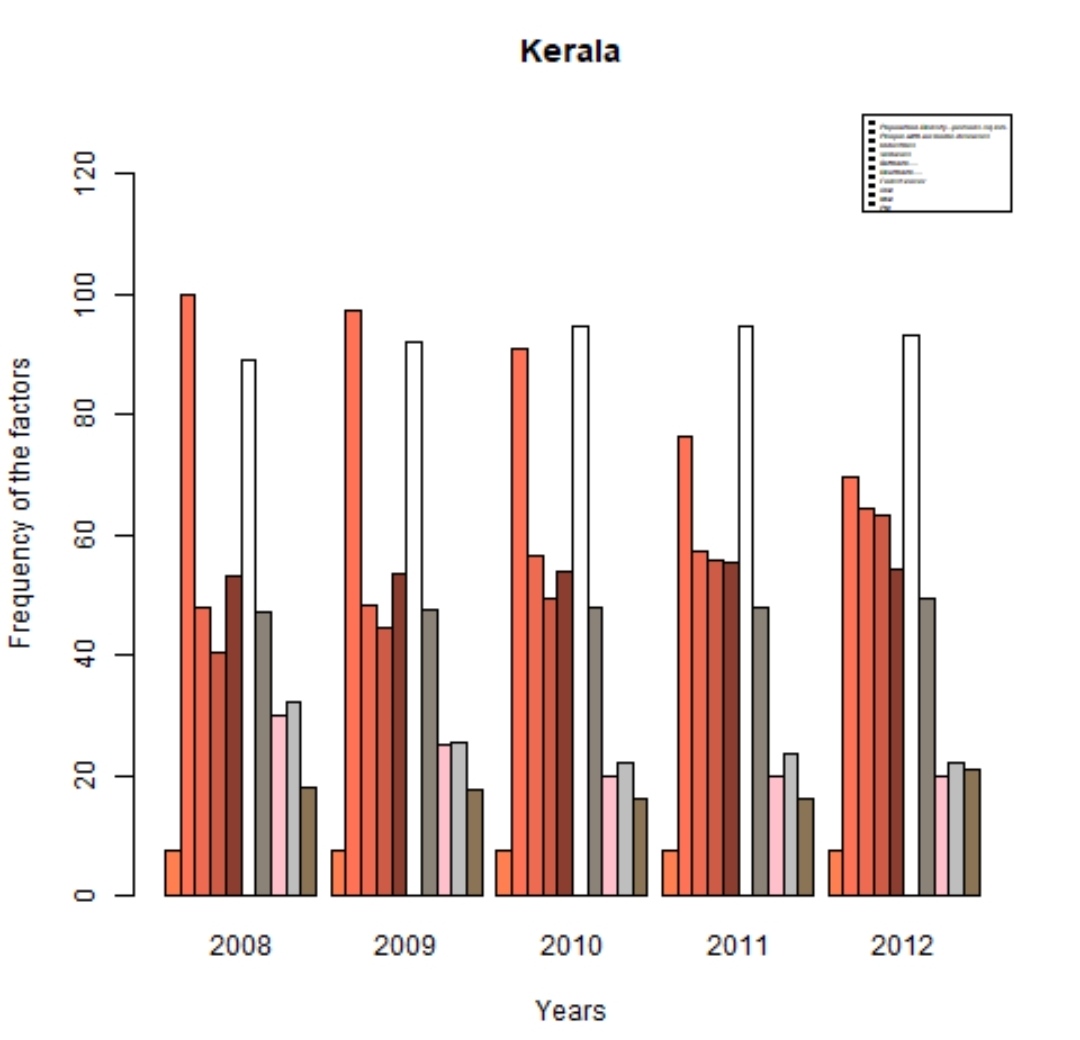


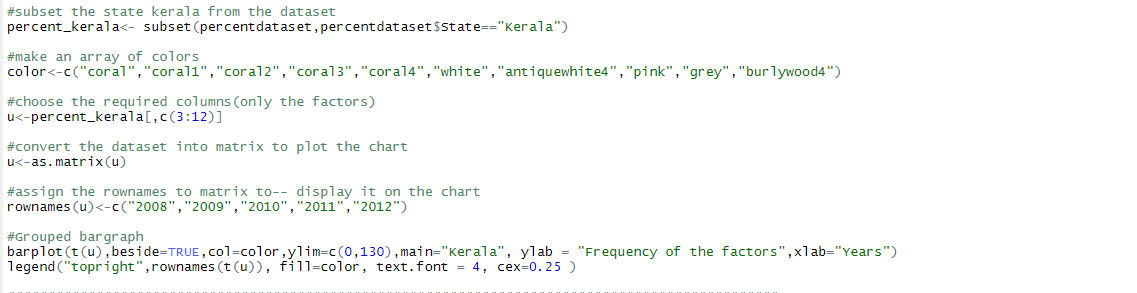
The following grouped bar graphs show the variation in frequency of each factor during the period of 2008-2012 for a particular state.

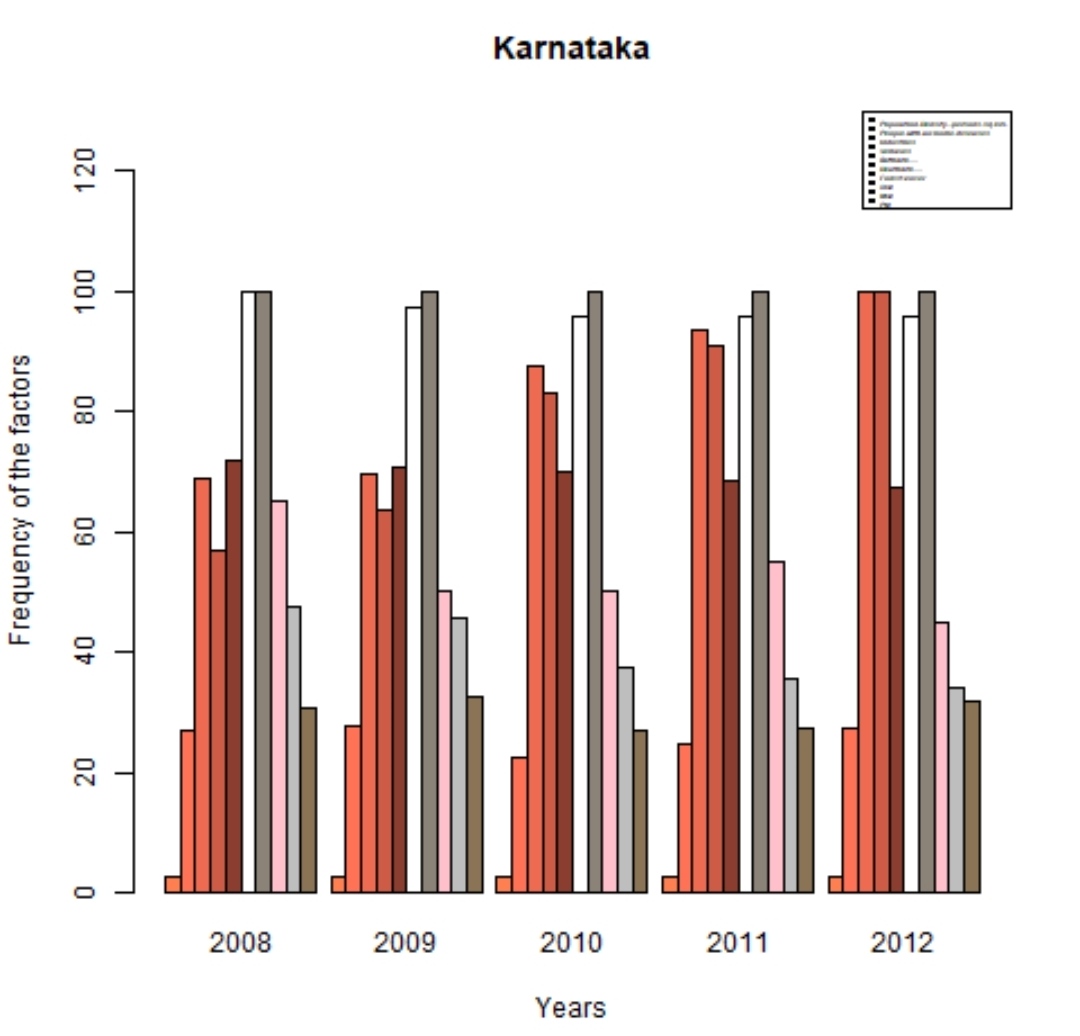


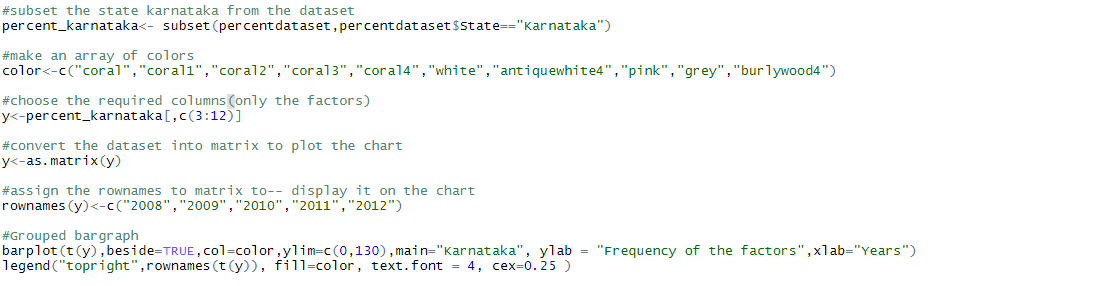


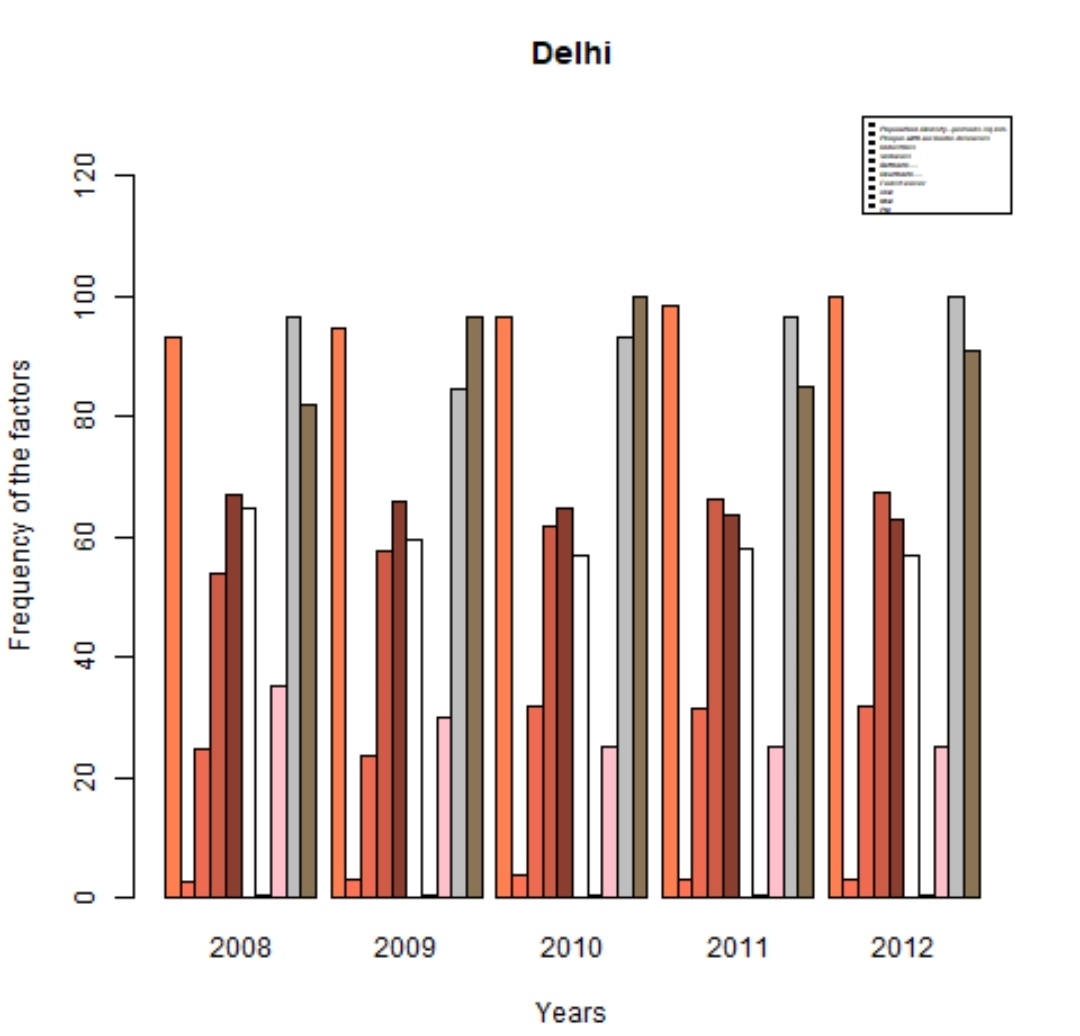


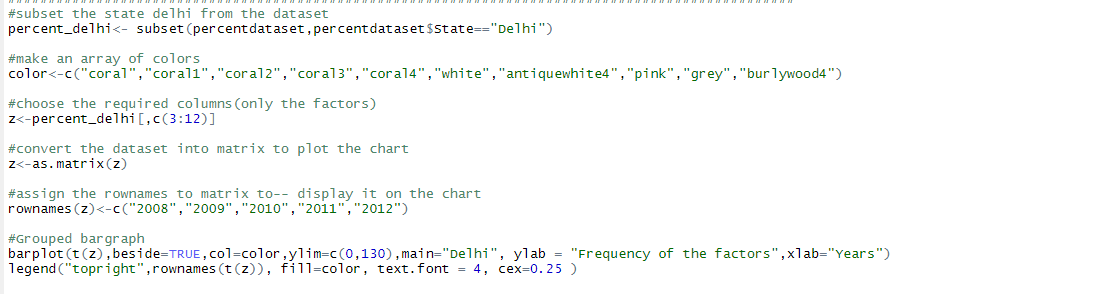




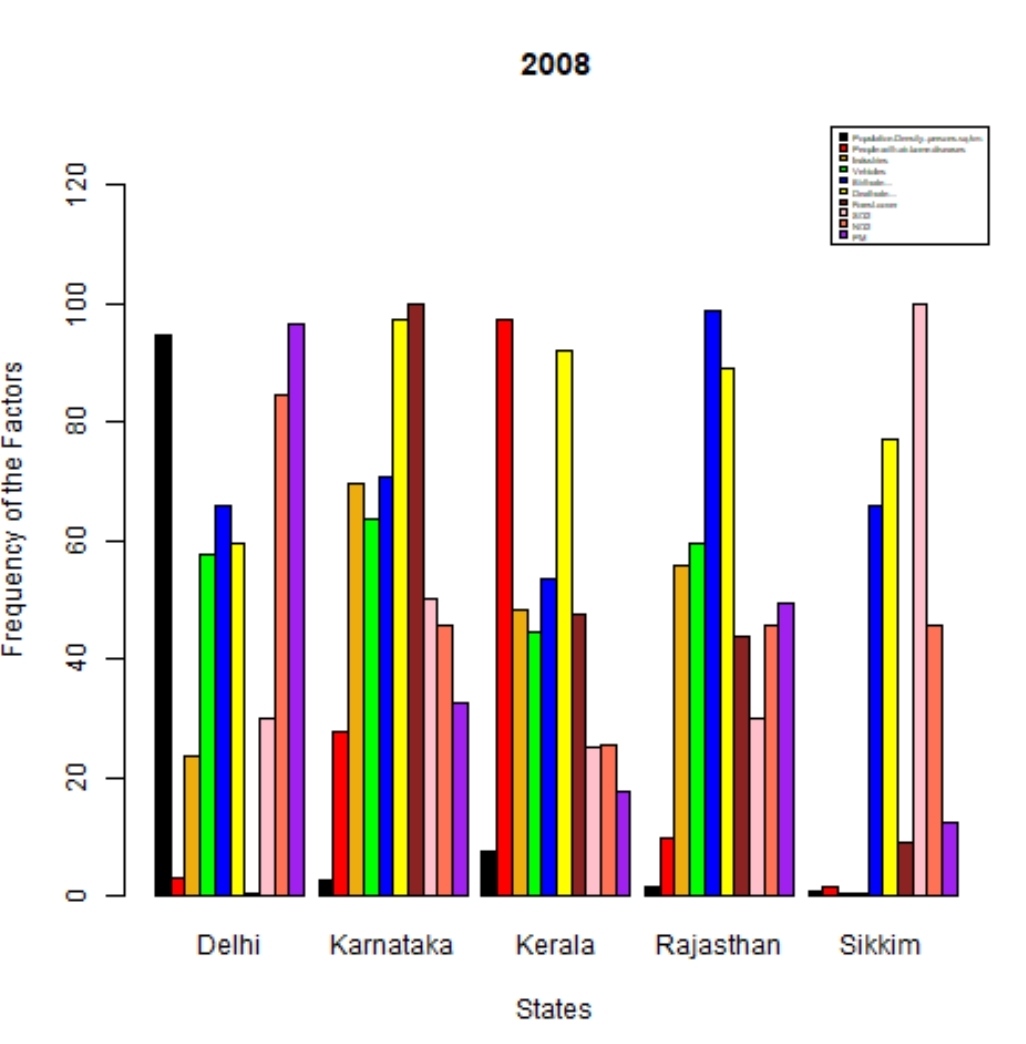


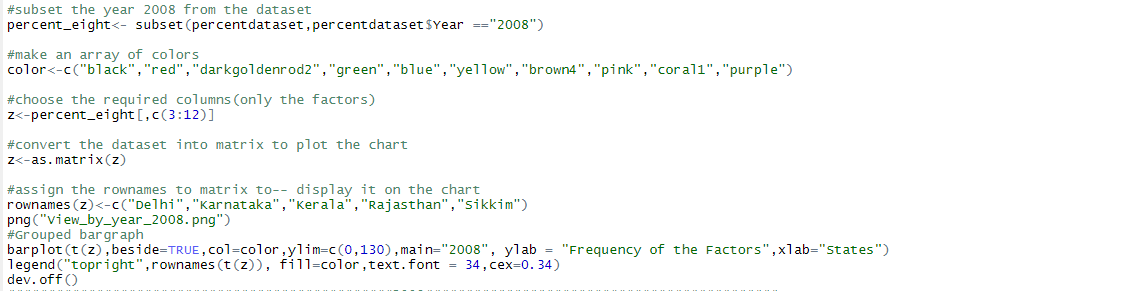


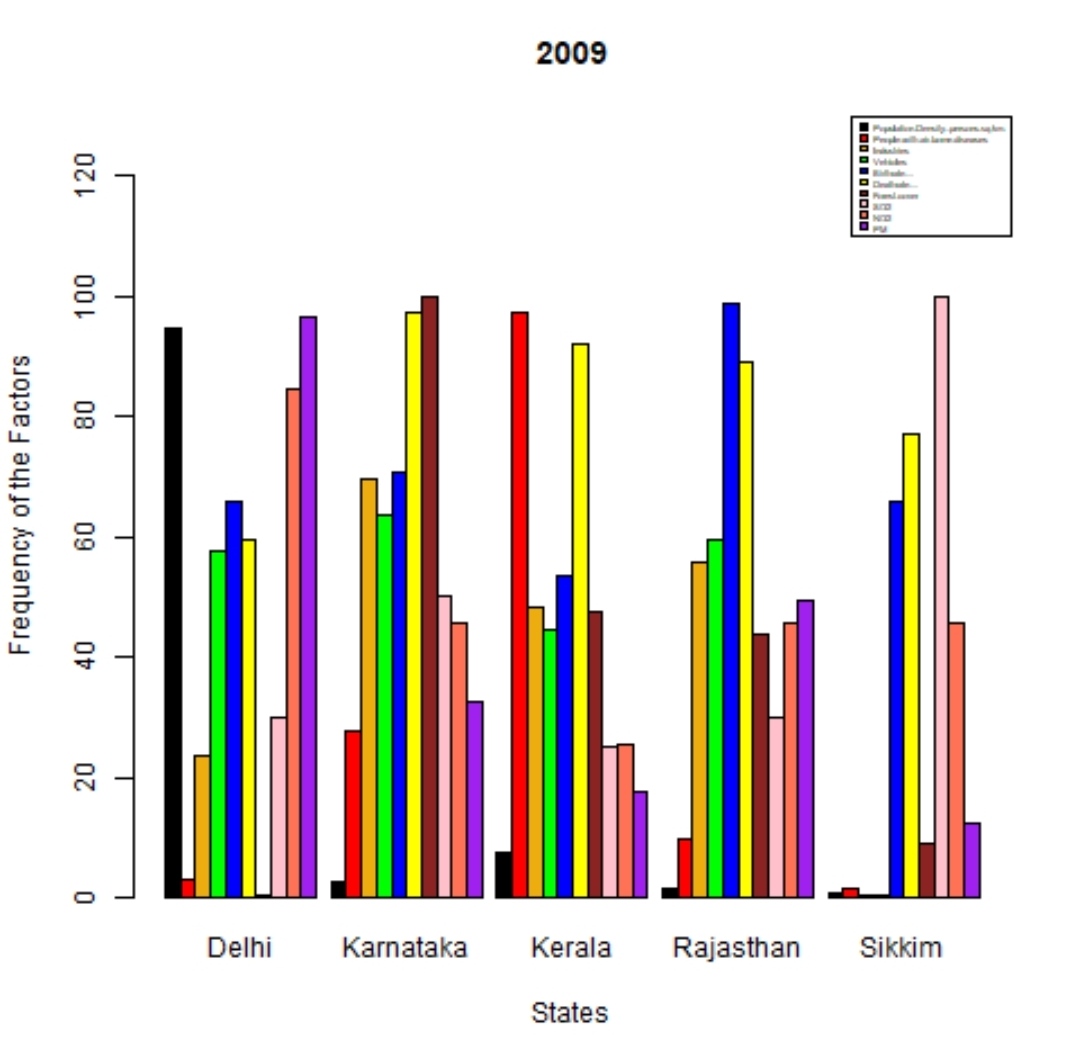


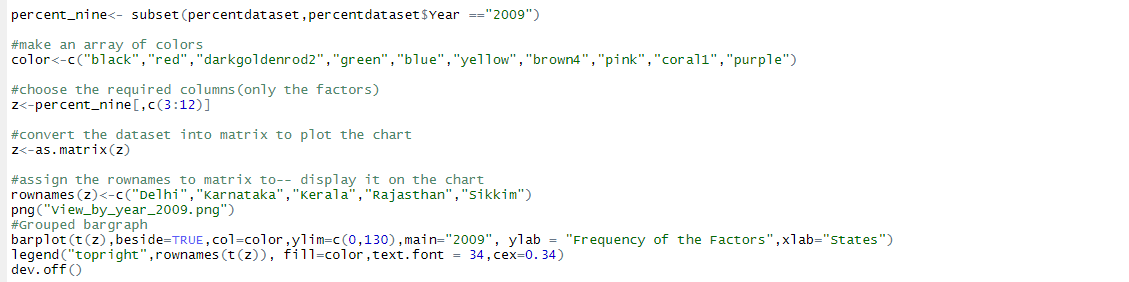


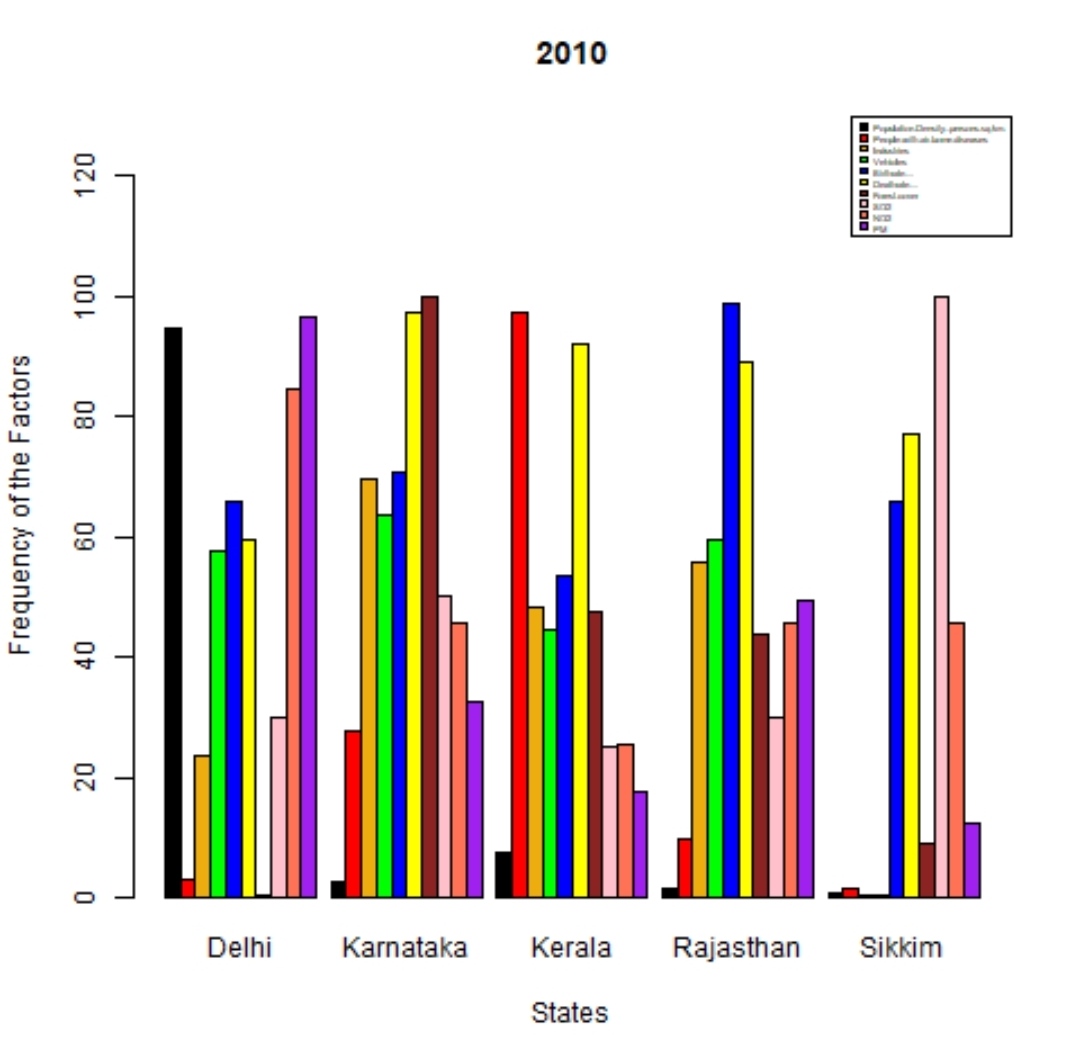
The following grouped bar graphs show the variation in frequency of each factor for all five states during a particular year.

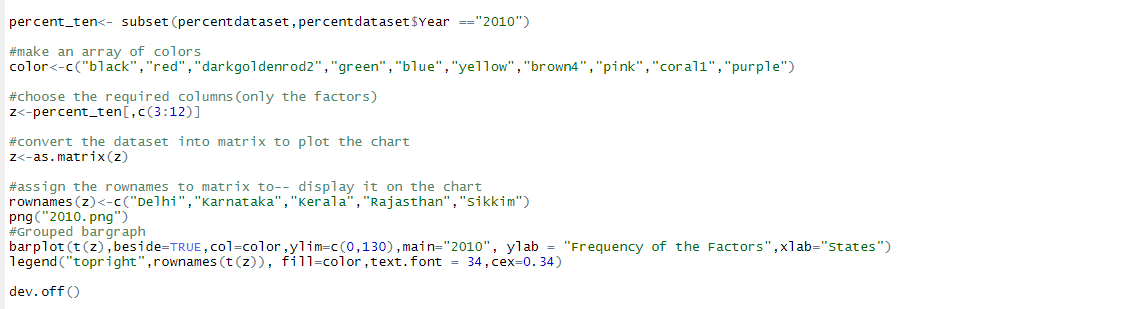


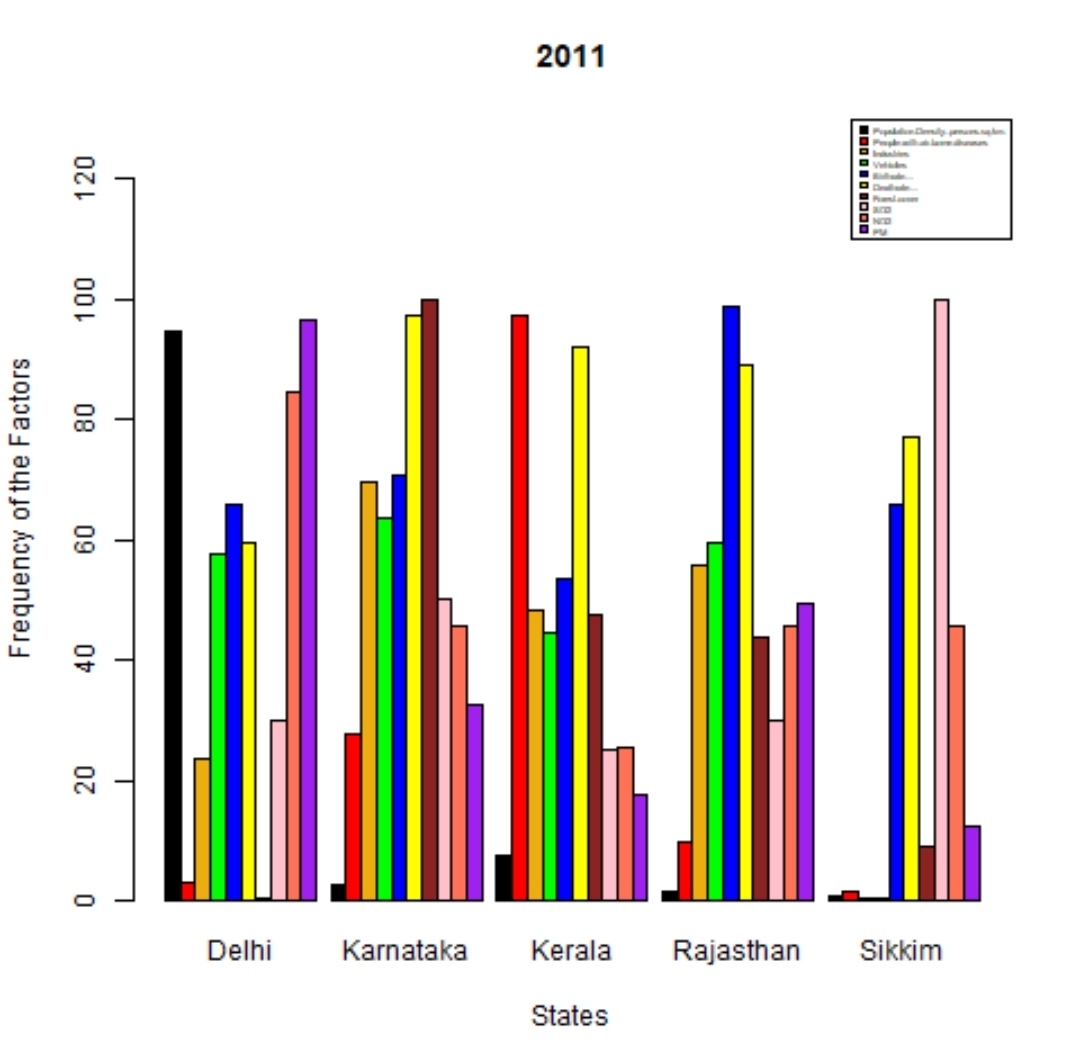




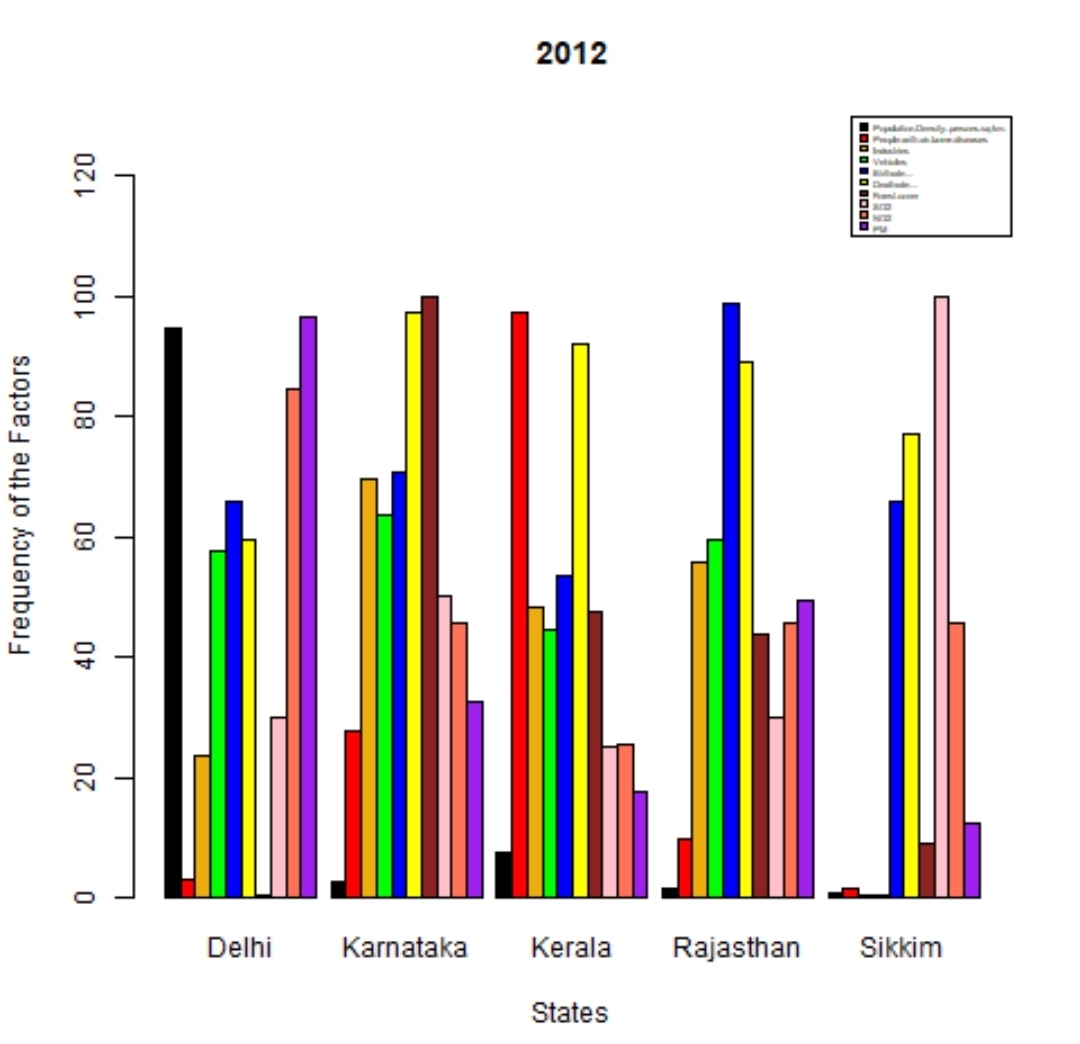


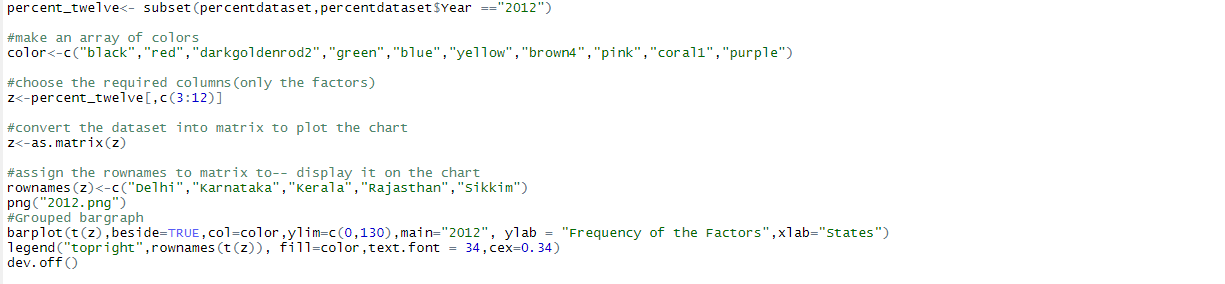




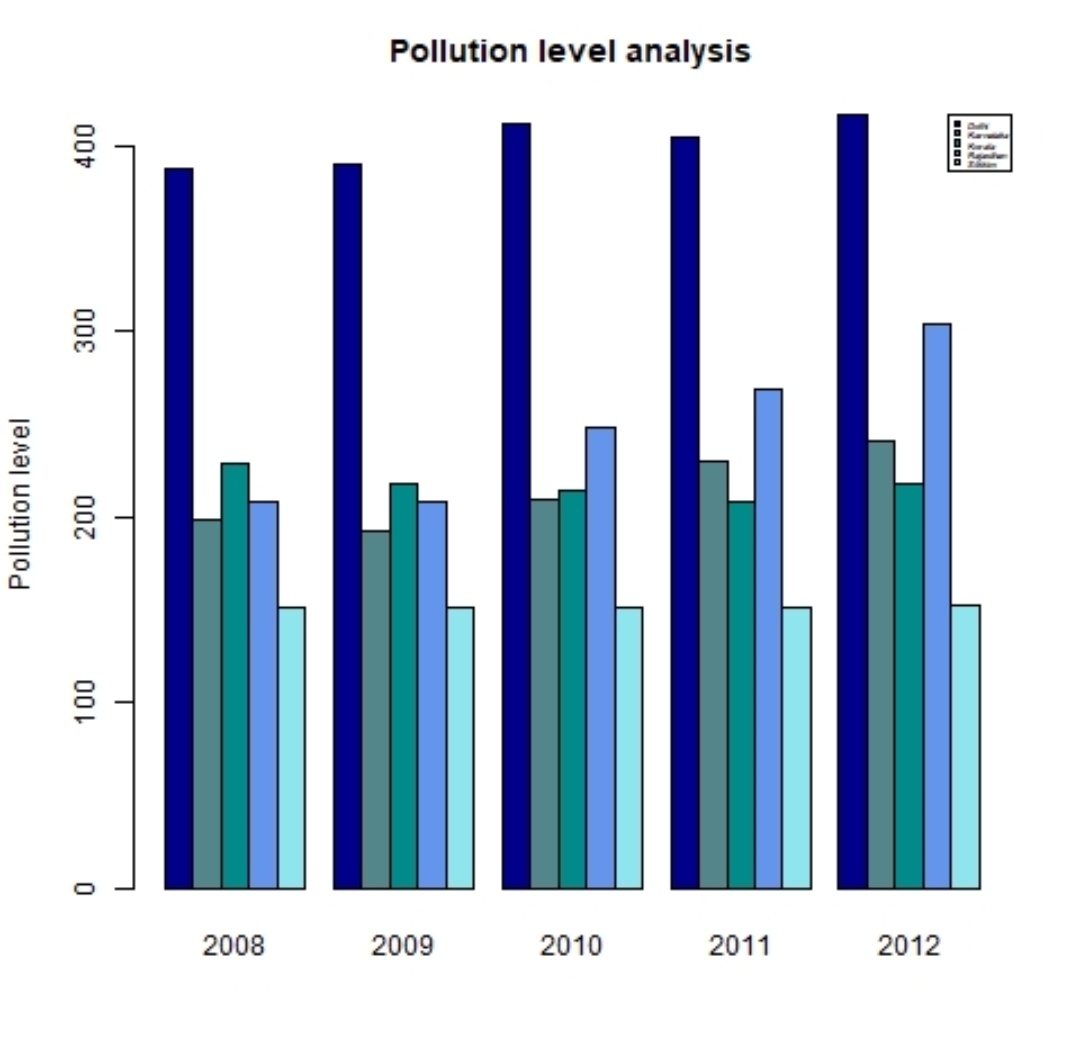


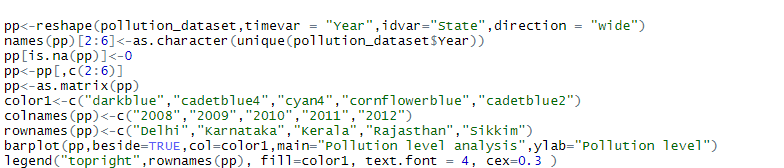




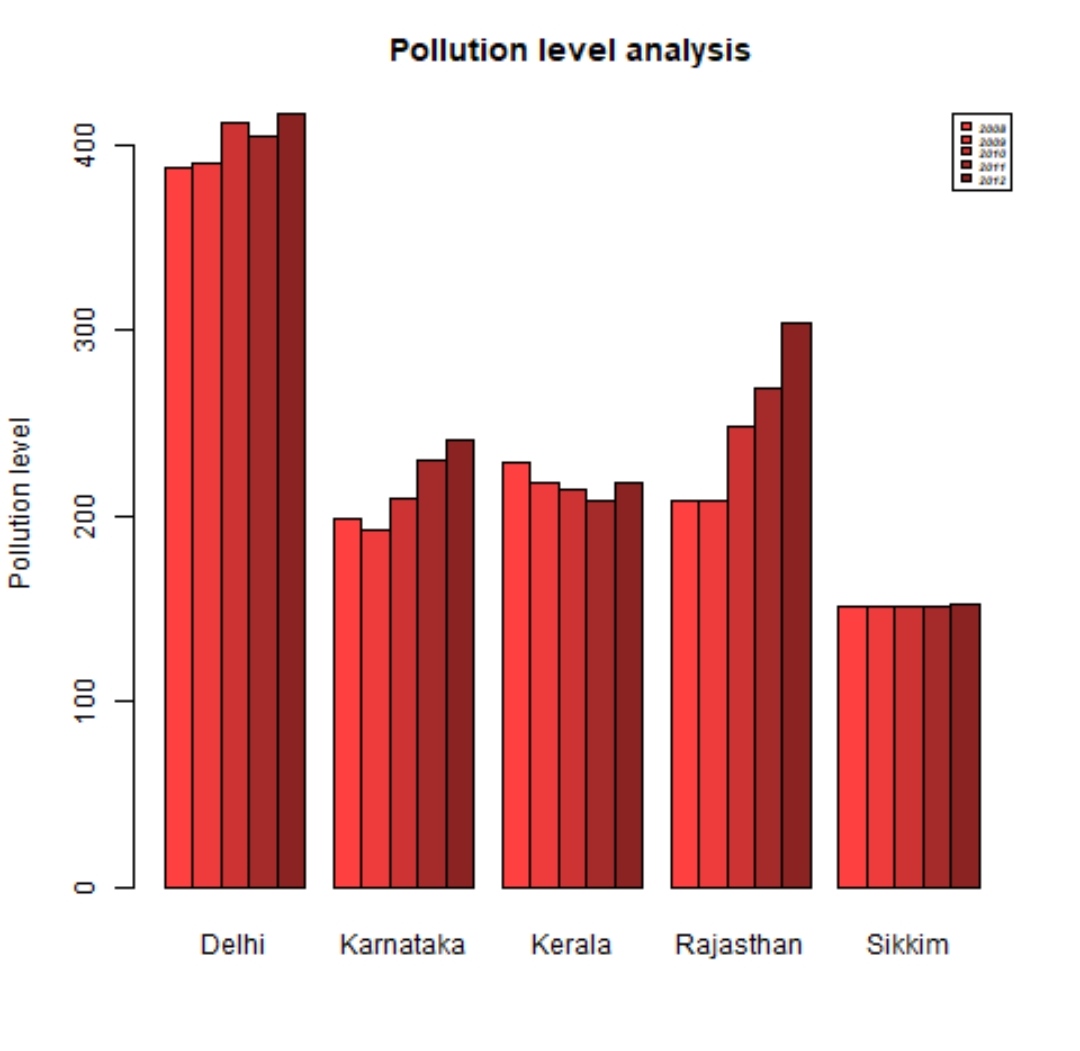


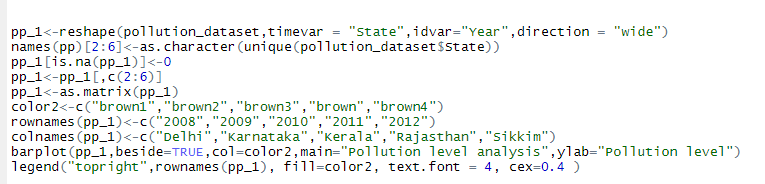
The below grouped bar graph shows the variation of pollution level in all five states during the period of 2008-2012.



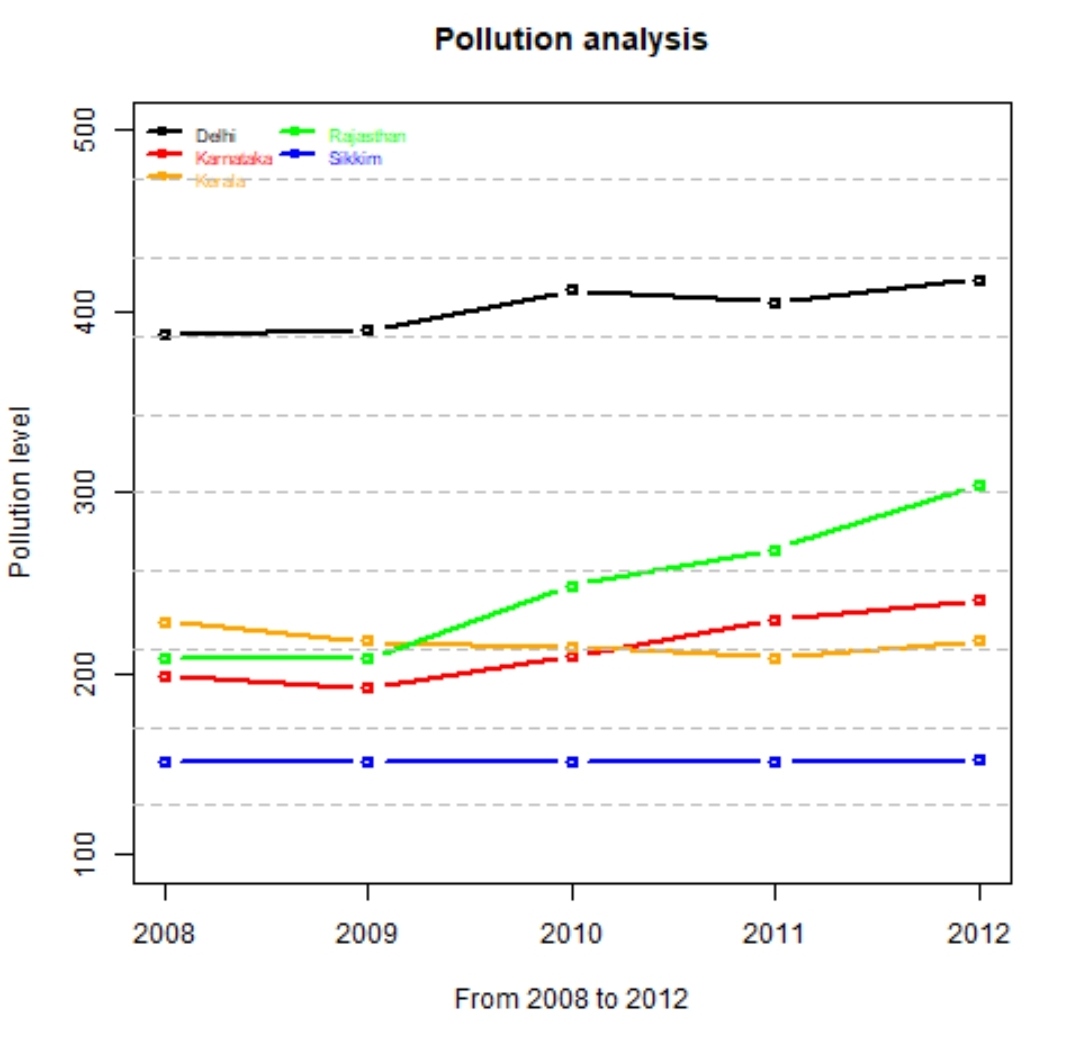


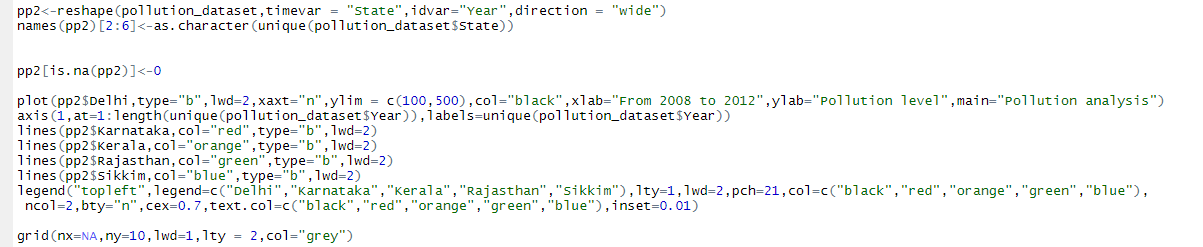
The below grouped bar graph shows the variation of pollution level during the period of 2008-2012 in all five states.





The below line graph shows the variation of pollution level in all five states during the period 2008-2012.





**CONCLUSION**

From the above analysis, we conclude that:

* The majorly affected state in India during the years 2008-2012 by air pollution is mostly in the northern region i.e. Delhi, which is heavily polluted and requires immediate action.
* Karnataka had initially lower levels of pollutants in the atmosphere (2008-2009) compared to the other states (Delhi, Karnataka and Kerala); however saw a steady increase of the same in the later years (2010-2012). This is largely due to the increase in population density per square km, industries, vehicles, and the concentration of pollutants in the air.
* Kerala was moderately polluted in the early stages (2008) but, later, were taken care of (2009-2011). The reason for the decrease could be awareness in citizens and government policies.
* Rajasthan saw a massive hike in the air pollution level during 2008-2012 which is directly correlated to the increase in population density per square km, industries, vehicles and particulate matter in the atmosphere.
* The pollution level in Sikkim has been constant throughout the years (2008-2011), however the levels increased by a very small factor in 2012 which can be attributed to the steady increase in population density, industry, vehicles and other determinants.

From the above data analysis approach, we conclude that data analysis is a crucial aspect for a better future. It is interesting to see how data analysis and the day to day instances are coherent and how data analysis can be used to deal with significant problems.

We must find a cure to this significant problem as it is killing our nation slowly.

**REFERENCES**

* <https://towardsdatascience.com/india-air-pollution-data-analysis-bd7dbfe93841>
* <https://www.researchgate.net/publication/320707293_Air_Quality_Prediction_Big_data_and_Machine_Learning_Approaches>
* <https://www3.epa.gov/airnow/2018conference/Plenary/karin-tuxen-bettman.pdf>
* <https://www.intel.co.uk/content/www/uk/en/it-management/cloud-analytic-hub/fighting-air-pollution.html>
* <https://www.sciencedirect.com/science/article/pii/S1877050918307555>
* <https://www.researchtrend.net/ijet/pdf/39-%20178.pdf>
* <https://www.analyticsvidhya.com/blog/2016/10/complete-study-of-factors-contributing-to-air-pollution/>