KSCHOOL | DATA SCIENCE MASTER | JULY 2021

MARCOS MARISCAL GARCÍA

HEALTH IS IN AIR

MARCOS MARISCAL GARCÍAr

[INTRODUCTION 3](#_Toc76753191)

[Preface 3](#_Toc76753192)

[Motivation 3](#_Toc76753193)

[Lessons from History 4](#_Toc76753194)

[About Air Quality data 5](#_Toc76753195)

[Interpreting the AQI 6](#_Toc76753196)

[Which Pollutants Can I Monitor Using the AQI? 7](#_Toc76753197)

[TARGET 9](#_Toc76753198)

[DATA SCIENCE PROCESS 10](#_Toc76753199)

[Step 1: Set a research goal 10](#_Toc76753200)

[Step 2: Collect the raw data needed for the study 10](#_Toc76753201)

[Step 3: Process the data for analysis 10](#_Toc76753202)

[Step 4: Explore the data 10](#_Toc76753203)

[Step 5: data modeling 10](#_Toc76753204)

[Step 6: Communicate results of the analysis 10](#_Toc76753205)

[DATA COLLECT THE RAW DATA NEEDED 11](#_Toc76753206)

[HOSPITAL VISITS & URGENCIES 11](#_Toc76753207)

[AIR QUALITY DATA 15](#_Toc76753208)

[POPULATION AND GEOGRAPHICAL DATA 17](#_Toc76753209)

[PROCESS THE DATA FOR ANALYSIS 18](#_Toc76753210)

# INTRODUCTION

### Preface

Last December 2020, a historic decision of the British justice recognized that Ella Adoo, a nine years girl, had a serious asthma attack causing her death because of a high level of pollution in South London in 2014

Delivering the narrative verdict, the warrant said that levels of nitrogen dioxide (NO2) near Ella's home exceeded World Health Organization and European Union guidelines.

Giving his conclusion over almost an hour, the coroner said: "I will conclude that Ella died of asthma, contributed to by exposure to excessive air pollution."

Las year World Health organization gives a terrifying death toll; air pollution kills an estimated seven million people worldwide every year.

From smog hanging over cities to smoke inside the home, air pollution poses a major [threat to health](https://www.who.int/airpollution/ambient/health-impacts/en/) and climate. The combined effects of ambient (outdoor) and household air pollution cause about seven million premature deaths every year, largely as a result of increased mortality from stroke, heart disease, chronic obstructive pulmonary disease, lung cancer and acute respiratory infections.

### Motivation

Since a member of my family had been diagnosed asthma, I am really concern about effects of Air pollution in human’s health. Our vacation destinations are always oriented to places with a good air quality. With this ever-present concern, we have learned to distinguish all the factors to take into account to know the quality of the air in a given place and time.

### Lessons from History

#### Donora. 1948

In October 1948, Donora, Pennsylvania, was enveloped in a lethal haze.

Over five days, nearly half of the town's 14,000 residents experienced severe respiratory or cardiovascular problems. It was difficult to breathe. The death toll rose to nearly 40.

Disturbing photos show Donora's streets hidden under a thick blanket of gray smog. A warm air pocket had passed high above the town, trapping cooler air below and sealing in pollutants.

|  |  |
| --- | --- |
| Nieblas asesinas, ¿fenómeno natural o paranormal? | La niebla tóxica de Donora de 1948 - el blog insostenible |

The situation in Donora was extreme, but it reflected a trend. Air pollution had become a harsh consequence of industrial growth across the country and world.

#### Meuse River valley. 1930

Belgium’s Meuse River valley has long been an inviting tourist destination. From Belgium it then crosses into Holland, where its name changes to the Maas River.

But for almost five days starting December 1, 1930, this scenic valley was occupied by a creeping horror — a wet, impenetrable and extremely toxic fog, centered at the town of Engis. Before it lifted completely on December 6 1930, 65 residents of Engis and several other towns, who had been exposed to the dense clouds, had died horrible, choking deaths. Hundreds of others were stricken with mysterious respiratory illnesses, and thousands of cattle, horses and other farm livestock asphyxiated and died in the fields “like flies sprayed with poison gas.”

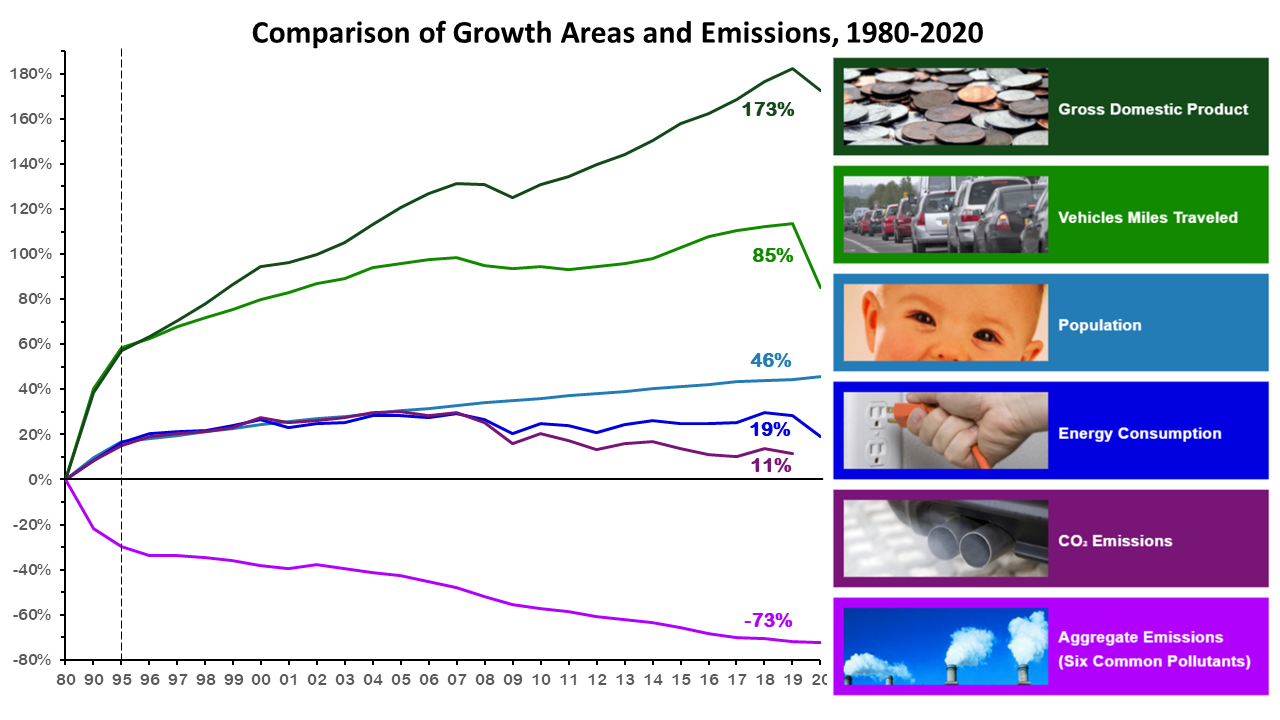
### About Air Quality data

The Air Quality Index, or AQI, was developed by the U.S. Environmental Protection Agency (EPA) to provide a simple, uniform way to report daily air quality conditions.

EPA creates air quality trends using measurements from monitors located across the country. The table below show that air quality based on concentrations of the common pollutants has improved nationally since 1980.

In 2020, about 68 million tons of pollution were emitted into the atmosphere in the United States. These emissions mostly contribute to the formation of ozone and particles, the deposition of acids, and visibility impairment.

The graph below shows that between 1980 and 2020, gross domestic product increased 173 percent, vehicle miles traveled increased 85 percent, energy consumption increased 19 percent, and U.S. population grew by 46 percent. During the same time period, total emissions of the six principal air pollutants dropped by 73 percent. The graph also shows that between 1980 and 2019, CO2 emissions increased by 11 percent.



### Interpreting the AQI

Air Quality Index

Think of the AQI as a yardstick that runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. For example, an AQI value of 50 or below represents good air quality, while an AQI value over 300 represents hazardous air quality.

The AQI is divided into six categories. Each category corresponds to a different level of health concern. Each category also has a specific color. The color makes it easy for people to quickly determine whether air quality is reaching unhealthy levels in their communities

| **Daily AQI Color** | | **Levels of Concern** | **Values of Index** | **Description of Air Quality** |
| --- | --- | --- | --- | --- |
| **Green** | **Good** | | **0 to 50** | **Air quality is satisfactory, and air pollution poses little or no risk.** |
| **Yellow** | **Moderate** | | **51 to 100** | **Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.** |
| **Orange** | **Unhealthy for Sensitive Groups** | | **101 to 150** | **Members of sensitive groups may experience health effects. The general public is less likely to be affected.** |
| **Red** | **Unhealthy** | | **151 to 200** | **Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.** |
| **Purple** | **Very Unhealthy** | | **201 to 300** | **Health alert: The risk of health effects is increased for everyone.** |
| **Maroon** | **Hazardous** | | **301 and higher** | **Health warning of emergency conditions: everyone is more likely to be affected.** |

### Which Pollutants Can I Monitor Using the AQI?

AQI numbers are determined by hourly measurements of five pollutants:

* fine particles (PM2.5),
* ground-level ozone (O3)
* sulfur dioxide (SO2)
* nitrogen dioxide (NO2)
* carbon monoxide (CO)

To check the de aka composite AQI, you need to take de maximum of all individual AQI:

AQI = max( AQIPM2.5, AQIPM10, AQIO3, ...)

The pollutant with the highest AQI value determines the overall AQI for that hour.

The five pollutants measured for the AQI are good indicators of daily air quality, but are not the only air pollutants, which may cause health effects, such as air toxics pollutants. Additionally, the AQI does not account for temperature or pollen levels, which may increase sensitivity to air pollutants.

The table below briefly describes each pollutant that goes into the AQI.

|  |  |  |
| --- | --- | --- |
| Pollutant | Abbreviation | Description |
| Ozone | O3 | Ozone is a form of oxygen with three atoms instead of the usual two atoms. It is a photochemical oxidant and, at ground level, is the main component of smog. Unlike other gaseous pollutants, ozone is not emitted directly into the atmosphere. Instead, it is created in the atmosphere by the action of sunlight on volatile organic compounds and nitrogen oxides.  In general, higher levels of ozone usually occur on sunny days with light winds, primarily from March through October. An ozone exceedance day is counted if the measured eight-hour average ozone concentration exceeds the standards. |
| Carbon Monoxide | CO | Carbon monoxide is a colorless, odorless, very toxic gas produced by the incomplete combustion of carbon-containing fuels, most notably by gasoline powered engines, power plants, and wood fires.  The eight-hour standard can be exceeded during winter months when very stable atmospheric conditions exist. |
| Sulfur Dioxide | SO2 | Sulfur dioxide is produced by burning sulfur-containing fuels (such as coal), smelting metallic ores containing sulfur, and removing sulfur from fuels. There are three sulfur dioxide standards which include a 24-hour average, an annual average, and a three-hour average. |
| Nitrogen Dioxide | NO2 | There are several oxides of nitrogen produced by high-temperature combustion. However, the National Ambient Air Quality Standard is only for Nitrogen Dioxide, which has an annual and 1-hour standard. |
| Particulate Matter | PM-2.5 PM-10 | Particle pollution (also called particulate matter or PM) is the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small, they can only be detected using an electron microscope. Particle pollution includes inhalable coarse particles, with diameters larger than 2.5 micrometers and smaller than 10 micrometers and fine particles, with diameters that are 2.5 micrometers and smaller. How small is 2.5 micrometers? Think about a single hair from your head. The average human hair is about 70 micrometers in diameter -- making it 30 times larger than the largest fine particle. These particles come in many sizes and shapes and can be made up of hundreds of different chemicals. Some particles, known as primary particles, are emitted directly from a source, such as construction sites, unpaved roads, fields, smokestacks or fires. Others form in complicated reactions in the atmosphere of chemicals such as sulfur dioxides and nitrogen oxides that are emitted from power plants, industries and automobiles. These particles, known as secondary particles, make up most of the fine particle pollution in the country.  Coarse particulates (PM-10) come from sources such as windblown dust from the desert or agricultural fields (sand storms) and dust kicked up on unpaved roads by vehicle traffic. PM-10 data is the near real-time measurement of particulate matter 10 microns or less in size from the surrounding air. This measurement is made at standard conditions, meaning it is corrected for local temperature and pressure.  Fine particulates (PM-2.5) are generally emitted from activities such as industrial and residential combustion and from vehicle exhaust. Fine particles are also formed in the atmosphere when gases such as sulfur dioxide, nitrogen oxides, and volatile organic compounds, emitted by combustion activities, are transformed by chemical reactions in the air. Large-scale agricultural burning or sand storms can produce huge volumes of fine particulates. PM-2.5 data is the near real-time measurement of particulate matter 2.5 microns or less in size from the surrounding air. This measurement is made at local conditions, and is not corrected for temperature or pressure. |

# TARGET

My Purpose is to analyze Air Quality data and cross it with the rates of emergency hospital visit related to asthma and other respiratory difficulties.

The idea is to recognize patterns between high levels of pollution in some geographical areas and the emergency hospital visits because of asthma and other breathings difficulties in the analyzed area.

The final target is to generate a predictive model in order to forecast the average hospital visits in future periods regarding air quality conditions.

# DATA SCIENCE PROCESS

We can consider **five** steps in the Data Science process, from when you give shape to an idea and after all when you present the results

## Step 1: Set a research goal

We have already defined our goal in the previous section. The goal is to recognize patterns between high levels of pollution in some geographical areas and the emergency hospital visits because of asthma and other breathings difficulties in the analyzed area.

Finally, we should be able to generate a predictive model to get future values of Hospital Visits rates with [respiratory diseases](https://www.linguee.es/ingles-espanol/traduccion/respiratory+diseases.html) according to Air Quality indexes in the same geographical area.

## Step 2: Collect the raw data needed for the study

In this case, we need to find at least three dataset in order to construct the database on which to carry out the study.

* HOSPITAL VISITS & URGENCIES
* AIR QUALITY DATA
* POPULATION AND GEOGRAPHICAL DATA

## Step 3: Process the data for analysis

Now that we have the raw data, it is time to prepare it. This includes transforming the data from a raw form into data that is directly usable in our models.

## Step 4: Explore the data

The goal of this step is to gain a deep understanding of the data. We will look for patterns, correlations, and deviations based on visual and descriptive techniques

## Step 5: data modeling

It is now that we attempt to gain the insights or make the predictions stated in our project objectives

## Step 6: Communicate results of the analysis

The last step of the data science model is presenting our results

# DATA COLLECT THE RAW DATA NEEDED

After many searches for information on the four data sets required, I finally decided to focus on the country with the largest number of public websites with data repositories about air quality data and Hospital visits & urgencies. I refer to US and the government sites for this kind of information.

### HOSPITAL VISITS & URGENCIES

<https://www.hcup-us.ahrq.gov>

The HCUP family of healthcare databases and related software tools and products has been made possible by a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality (AHRQ).

The HCUP Summary Trend Tables include monthly information on hospital utilization derived from the HCUP State Inpatient Databases and HCUP State Emergency Department Databases.

#### The Data Set specifications

|  |  |
| --- | --- |
| **Patient/Stay Characteristics** | **Description** |
| **Month** |  |
| All discharges | Based on discharges in the month. For Florida, trends are reported on a quarterly basis. |
| **Age in Years at Admission** | Age was imputed if missing. For records missing age, the age was assigned the average age for the MS-DRG (calculated within MS-DRG version). Age is rarely missing in the HCUP databases. Only 342 of the 7.2 million records in the 2017 HCUP National Inpatient Sample (NIS) are missing age.   The age groups reported vary across the HCUP Summary Trend Tables. Information on age is not available for the following: encounter type of normal newborns, encounter type of deliveries, and service line for maternal/neonatal conditions. |
| **Sex** | Sex was imputed if missing. For records missing sex, the sex was assigned based on the proportion of cases for the MS-DRG (calculated within MS-DRG version). Sex is rarely missing in the HCUP databases. Only 950 of the 7.2 million records in the 2017 HCUP National Inpatient Sample (NIS) are missing sex.   Information on sex is not available for encounter type of deliveries. |
| Male |  |
| Female |  |
| **Patient Race/Ethnicity** |  |
| White, nonHispanic |  |
| Black, nonHispanic |  |
| Hispanic |  |
| All other nonHispanics including Asian/Pacific Islander and Native American/Alaskan Native |  |
| Race/ethnicity data not included in the SID | Not all SID include information on patient race/ethnicity. This value is used in the HCUP Summary Trend Tables when all records in the SID are missing race/ethnicity. |
| Race/ethnicity data included in the SID but missing |  |
| **Urban-Rural Location of Patient's Residence** | The ZIP Code of the patient's residence is assigned to a county based on the geographic centroid. The urban-rural assignment is based on National Center for Health Statistics (NCHS) designation for the county. |
| Large metro |  |
| Medium and small metros |  |
| Nonmetro |  |
| Missing | Information is missing from the HCUP Summary Trend Tables when any of the following occurs: patient’s ZIP Code is invalid, patient’s ZIP Code is missing, patient’s ZIP Code indicates the patient is foreign. |
| **Community Income** | Quartiles are defined so that the total U.S. population is evenly distributed. Cut-offs for the quartiles are determined annually using ZIP Code demographic data obtained from Claritas, a vendor that produces population estimates and projections. The quartile assignment is based on the median household income of the ZIP Code of the patient's residence. This information was obtained from Claritas. |
| Quartile 1, lowest income |  |
| Quartile 2-3, middle income quartiles |  |
| Quartile 4, highest income |  |
| Missing | Information is missing in the HCUP Summary Trend Tables when any of the following occurs: patient’s ZIP Code is invalid, patient’s ZIP Code is missing, patient’s ZIP Code indicates the patient is foreign. |
| **Expected Primary Payer** |  |
| Medicare | Medicare includes fee-for-service and managed care Medicare. |
| Medicaid | Medicaid includes fee-for-service and managed care Medicaid. |
| Private insurance | Private insurance includes commercial nongovernmental payers, regardless of the type of plan (e.g., private health maintenance organizations, preferred provider organizations). |
| Self-pay/no charge | Self-pay/No charge includes self-pay, no charge, charity, and no expected payment. |
| Other | Other payers include other Federal and local government programs (e.g., TRICARE, CHAMPVA, Indian Health Service, Black Lung, Title V) and Workers' Compensation. |
| Missing |  |
| **Urban-Rural Location of the Hospital** | The ZIP Code of the hospital is assigned to a county based on the geographic centroid. The urban-rural assignment is based on National Center for Health Statistics (NCHS) designation for the county. |
| Large metro |  |
| Medium and small metros |  |
| Nonmetro |  |
| **Procedure Class** | Information on operating room procedures is not available for the following: encounter type of normal newborn and service line of maternal/neonatal conditions. |
| At least one operating room surgery during the inpatient stay | Identification of an operating room surgical procedure is based on the Procedure Classes Refined for ICD-10-PCS, v2021.2. At least one ICD-10-PCS code must be identified as a major therapeutic or diagnostic procedure. If information on procedure day is missing on the SID record or the SID does not include procedure days, the operating room procedure is set to occur on the day of admission. This count will differ from count of surgical discharges under the hospital service line for two reasons: (1) it is not assigned hierarchically; (2) it is based on the procedure class, not MS-DRGs. Information on the Procedure Classes is available at www.hcup-us.ahrq.gov/toolssoftware/procedureicd10/procedure\_icd10.jsp. |
| At least one major therapeutic procedure during the inpatient stay | Identification is based on the Procedure Classes Refined for ICD-10-PCS, v2021.2. At least one ICD-10-PCS code must be identified as a major therapeutic (procedure class = 4). If information on procedure day is missing on the SID record or the SID does not include procedure days, the major therapeutic procedure is set to occur on the day of admission. Information on the Procedure Classes is available at www.hcup-us.ahrq.gov/toolssoftware/procedureicd10/procedure\_icd10.jsp. |
| At least one major diagnostic procedure during the inpatient stay | Identification is based on the Procedure Classes Refined for ICD-10-PCS, v2021.2. At least one ICD-10-PCS code must be identified as a major diagnostic (procedure class = 3). If information on procedure day is missing on the SID record or the SID does not include procedure days, the major diagnostic procedure is set to occur on the day of admission. Information on the Procedure Classes is available at www.hcup-us.ahrq.gov/toolssoftware/procedureicd10/procedure\_icd10.jsp. |
| **Type of Intensive Care Use** | Information on intensive care use is not available for the following: encounter type of normal newborn, encounter type of deliveries, and service line of maternal/neonatal conditions. |
| Any use of intensive care during the inpatient stay | Information is dependent on two data elements in the SID: revenue center codes which identify types of intensive care services and units of service associated with each revenue center which specifies the number of days the patient was in the intensive care. Not all SID include these data elements. Any use of intensive care includes intensive care (revenue center codes 0200-0202), cardiac intensive care (revenue center codes 0210-0219), neonatal intensive care (revenue center codes 0173-0174), pediatric intensive care (revenue center code 0203), psychiatric intensive care (revenue center code 0204), intermediate intensive care (revenue center code 0206), burn care (revenue center code 0207), trauma care (revenue center code 0208), and other intensive care (revenue center code 0209). |
| Any use of the intensive care unit (ICU) during the inpatient stay | Information is dependent on two data elements in the SID: revenue center codes which identify types of intensive care services and units of service associated with each revenue center which specifies the number of days the patient was in the intensive care. Not all SID include these data elements. Intensive care units are identified by revenue center codes 0200-0202. |
| Any use of the cardiac care unit (CCU) during the inpatient stay | Information is dependent on two data elements in the SID: revenue center codes which identify types of intensive care services and units of service associated with each revenue center which specifies the number of days the patient was in the intensive care. Not all SID include these data elements. Cardiac intensive care units are identified by revenue center codes 0210-0219. |
| Any use of the neonatal intensive care unit (NICU) or pediatric intensive care unit (PICU) during the inpatient stay | Information is dependent on two data elements in the SID: revenue center codes which identify types of intensive care services and units of service associated with each revenue center which specifies the number of days the patient was in the intensive care. Not all SID include these data elements. Neonatal intensive care units are identified by revenue center codes 0173-0174 and pediatric intensive care units are identified by revenue center code 0203. |
| Any use of other types of intensive care units (psychiatric, burn, trauma, etc.) during the inpatient stay | Information is dependent on two data elements in the SID: revenue center codes which identify types of intensive care services and units of service associated with each revenue center which specifies the number of days the patient was in the intensive care. Not all SID include these data elements. Other intensive care includes psychiatric intensive care (revenue center code 0204), intermediate intensive care (revenue center code 0206), burn care (revenue center code 0207), trauma care (revenue center code 0208), and other intensive care (revenue center code 0209). |
| **Mechanical Ventilation Use** | Information on mechanical ventilation is not available for the following: encounter type of normal newborn, encounter type of deliveries, and service line of maternal/neonatal conditions. |

### AIR QUALITY DATA

<https://aqs.epa.gov>

When American Congress writes an environmental law, we implement it by writing regulations.

The Goal is Cleaner, Healthier Environment: Deliver a cleaner, safer, and healthier environment for all Americans and future generations by carrying out the Agency’s core mission.

#### The Data Set specifications

Each daily summary file contains data for every monitor (sampled parameter) in our database for each day. The daily summary files contain (at least) one record for each monitor that reported data for the given day

The file is comma separated variables (CSV) with a header row.

| **Field Position** | **Field Name** | **Description** |
| --- | --- | --- |
| 1 | State Code | The FIPS code of the state in which the monitor resides. |
| 2 | County Code | The FIPS code of the county in which the monitor resides. |
| 3 | Site Num | A unique number within the county identifying the site. |
| 4 | Parameter Code | The AQS code corresponding to the parameter measured by the monitor. |
| 5 | POC | This is the “Parameter Occurrence Code” used to distinguish different instruments that measure the same parameter at the same site. |
| 9 | Parameter Name | The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants. |
| 12 | Date Local | The calendar date for the summary. All daily summaries are for the local standard day (midnight to midnight) at the monitor. |
| 13 | Units of Measure | The unit of measure for the parameter. QAD always returns data in the standard units for the parameter. Submitters are allowed to report data in any unit and EPA converts to a standard unit so that we may use the data in calculations. |
| 14 | Event Type | Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data form them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor. |
| 15 | Observation Count | The number of observations (samples) taken during the day. |
| 16 | Observation Percent | The percent representing the number of observations taken with respect to the number scheduled to be taken during the day. This is only calculated for monitors where measurements are required (e.g., only certain parameters). |
| 17 | Arithmetic Mean | The average (arithmetic mean) value for the day. |
| 18 | 1st Max Value | The highest value for the day. |
| 19 | 1st Max Hour | The hour (on a 24-hour clock) when the highest value for the day (the previous field) was taken. |
| 20 | AQI | The Air Quality Index for the day for the pollutant, if applicable. |
| 29 | Date of Last Change | The date the last time any numeric values in this record were updated in the AQS data system. |

### POPULATION AND GEOGRAPHICAL DATA

https://www2.census.gov

The Census Bureau's mission is to serve as the nation’s leading provider of quality data about its people and economy

Our goal is to provide the best mix of timeliness, relevancy, quality and cost for the data we collect and services we provide

#### The Data Set specifications

|  |  |
| --- | --- |
| VARIABLE | DESCRIPTION |
| SUMLEV | Geographic summary level |
| REGION | Census Region code |
| DIVISION | Census Division code |
| STATE | State FIPS code |
| NAME | State name |
| CENSUS2010POP | 4/1/2010 resident total Census 2010 population |
| ESTIMATESBASE2010 | 4/1/2010 resident total population estimates base |
| POPESTIMATE2010 | 7/1/2010 resident total population estimate |
| POPESTIMATE2011 | 7/1/2011 resident total population estimate |
| POPESTIMATE2012 | 7/1/2012 resident total population estimate |
| POPESTIMATE2013 | 7/1/2013 resident total population estimate |
| POPESTIMATE2014 | 7/1/2014 resident total population estimate |
| POPESTIMATE2015 | 7/1/2015 resident total population estimate |
| POPESTIMATE2016 | 7/1/2016 resident total population estimate |
| POPESTIMATE2017 | 7/1/2017 resident total population estimate |
| POPESTIMATE2018 | 7/1/2018 resident total population estimate |
| POPESTIMATE2019 | 7/1/2019 resident total population estimate |
| POPESTIMATE2020 | 7/1/2020 resident total population estimate |

# PROCESS THE DATA FOR ANALYSIS

Here I describe followed steps for preprocesing the data set. I have done the whole preprocesing data in RMarkDown “DataCleansing.Rmd”

The best to follow all steps made in this R notebook is to attach the hmtl generated with the RMarkDown

# DataCleansing

#### Marcos Mariscal Garcia

#### 21/5/2021

This RMarkdown **loads**, **clean** and **prepare** data for further analysis

## Loading First Data Set: “Dialy Air Quality Monitor”

From: <https://www.hcup-us.ahrq.gov/reports/trendtables/summarytrendtables.jsp#export>.

This Data Set contains the daily summary files with one record for each monitor that reported data for the given day

After Loading, we aggregate AQI US County Data by Month and per Year in order to prepare this data to merge with “HOspital Visits & Emergencies” Data Set from 2017 to 2020

C:\Users\mmariscal\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\86F3BA42.tmp

A caption

Each NAAQS pollutant has a separate AQI scale, with an AQI rating of 100 corresponding to the concentration of the Federal Standard for that pollutant.

And this is the tibble with AQI Data aggregated by month:

head(aggtot)

## # A tibble: 6 x 8

## State.Name county.Name yearmm Cat Category AQI Def\_parm aggnum

## <chr> <chr> <date> <dbl> <chr> <int> <chr> <int>

## 1 Alabama Baldwin 2017-01-01 1 Good 30 PM2.5 1

## 2 Alabama Baldwin 2017-02-01 1 Good 38 PM2.5 1

## 3 Alabama Baldwin 2017-03-01 2 Moderate 58 PM2.5 2

## 4 Alabama Baldwin 2017-04-01 2 Moderate 74 Ozone 1

## 5 Alabama Baldwin 2017-05-01 3 Unhealthy for S~ 108 Ozone 1

## 6 Alabama Baldwin 2017-06-01 2 Moderate 65 PM2.5 1

### Merging with geographical data for US Counties, from r package (“sf”), and saving as RData

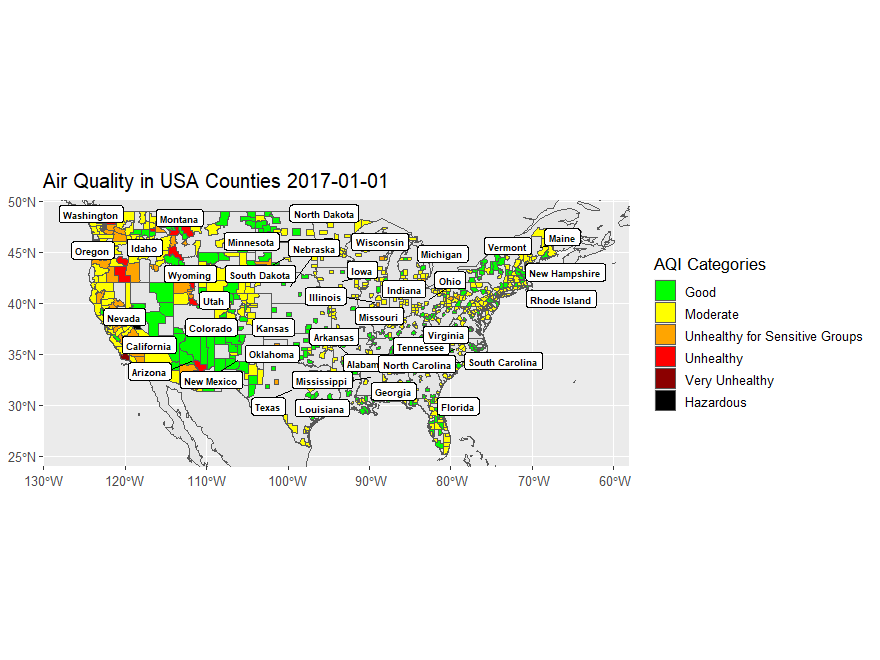
Next step is to merge **AQI** Data with US **Counties** geographical data with the objective to plot a US map with those values

### Representing Air Quality on a US geographical detailed by county

–

#### Animated Map with Air Quality Indexes by US counties

In the map below we can see that the data collected is really useful for our purposes. The map represent Air Quality Indexes For the 2017 to 2020 years:



## Loading Second Data Set “Hospital Visits & Emergencies”

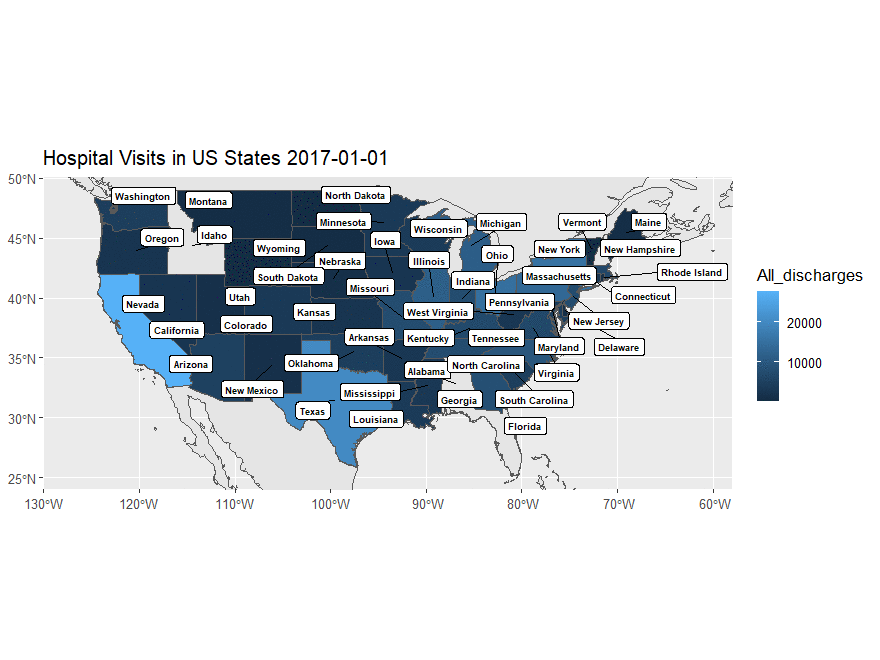
From: <https://www.hcup-us.ahrq.gov/reports/trendtables/summarytrendtables.jsp#export>

This is the data for the **Hospital** **Visits** and **Emergencies** in US States from **2017** to **2020**

This data is an excel report and the different States date are in different **sheets** of the same **excel** **workbook**

### Map for Hospital Visits by US State per year and month

We have merged geographical US states data with the Hospital visits dataset to generate the animated map below



## Loading the Third Data Set: US States population data from 2017 to 2020

From: <https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/state/totals/USA>

With this data we can calculate the rates for Hospital Visits & Emergencies per Population in each US State

## # A tibble: 6 x 6

## REGION DIVISION STATE NAME POPULATION YEAR

## <chr> <chr> <int> <chr> <int> <dbl>

## 1 0 0 0 United States 325122128 2017

## 2 0 0 0 United States 326838199 2018

## 3 0 0 0 United States 328329953 2019

## 4 0 0 0 United States 329484123 2020

## 5 1 0 0 Northeast Region 56083383 2017

## 6 1 0 0 Northeast Region 56084543 2018

#### We can see with these two animated maps that the **collected** **data** are **good** and **accuracy** for the goal of this project

# Exploratory\_Data\_Analysis

#### Marcos Mariscal

#### 23/5/2021

#### This RMarkdown **merge** all data and represent different graphs to visualice relation with features

### Loading Air QUality Indexes Data and Hospital Visits $ Emergencies population at State level with geographical data

This first DataSet has been generated with the DataCleasing R Notebook

## ID State.Name county.Name yearmm Cat Category AQI Def\_parm

## 1 alabama,baldwin Alabama Baldwin 2020-05-01 2 Moderate 74 Ozone

## 2 alabama,baldwin Alabama Baldwin 2020-01-01 1 Good 48 PM2.5

## 3 alabama,baldwin Alabama Baldwin 2020-06-01 2 Moderate 61 Ozone

## 4 alabama,baldwin Alabama Baldwin 2017-07-01 2 Moderate 67 PM2.5

## 5 alabama,baldwin Alabama Baldwin 2020-02-01 1 Good 31 PM2.5

## 6 alabama,baldwin Alabama Baldwin 2018-02-01 1 Good 31 PM2.5

## aggnum geom area

## 1 1 MULTIPOLYGON (((-87.93757 3... 4107738938

## 2 1 MULTIPOLYGON (((-87.93757 3... 4107738938

## 3 1 MULTIPOLYGON (((-87.93757 3... 4107738938

## 4 1 MULTIPOLYGON (((-87.93757 3... 4107738938

## 5 1 MULTIPOLYGON (((-87.93757 3... 4107738938

## 6 1 MULTIPOLYGON (((-87.93757 3... 4107738938

### Loading Hospital Visits & Emergencies and the whole population at USA State level with geographical data

This second DataSet has been also generated with the DataCleasing R Notebook

## STATES Dates All\_discharges Ages\_0\_4 Ages\_5\_9 Ages\_10\_17 Ages\_18\_44

## 1 Alaska 2017-01-01 359 77 NA NA 15

## 2 Alaska 2017-02-01 364 87 NA NA 22

## 3 Alaska 2017-03-01 414 108 NA NA 28

## 4 Alaska 2017-04-01 283 75 NA NA 20

## 5 Alaska 2017-05-01 288 53 NA NA 18

## 6 Alaska 2017-06-01 245 46 NA NA 26

## Ages\_45\_64 Ages\_65\_79 Ages\_80 Male Female White Black Hispanic

## 1 107 105 46 184 175 200 NA 15

## 2 82 106 56 178 186 212 NA 13

## 3 101 96 68 210 204 201 12 NA

## 4 65 82 38 148 135 129 NA NA

## 5 64 120 26 143 145 147 NA NA

## 6 60 73 36 118 127 133 NA NA

## All\_other\_races Race\_missing Resi\_L\_metro Resi\_M\_S\_metros Resi\_nonmetro

## 1 131 NA NA 250 107

## 2 121 NA NA 232 132

## 3 182 15 NA 224 190

## 4 145 NA NA 154 129

## 5 122 NA NA 175 111

## 6 93 NA NA 152 85

## Resi\_missing Quart1\_lowest\_income Quart2\_3\_middle\_income

## 1 NA NA 106

## 2 NA 24 116

## 3 NA 40 132

## 4 NA 40 77

## 5 NA 19 79

## 6 NA 20 75

## Quart4\_highest\_income Missing\_income Medicare Medicaid Private\_insurance

## 1 211 32 176 122 39

## 2 201 23 166 116 54

## 3 205 37 190 140 46

## 4 140 26 122 111 22

## 5 151 39 151 81 26

## 6 123 27 110 83 28

## Self\_pay Other\_pay Hosp\_M\_S\_metros Hosp\_Nonmetro Proc\_operating\_room

## 1 NA 17 281 78 23

## 2 NA 20 263 101 19

## 3 11 27 282 132 21

## 4 NA 25 197 86 17

## 5 NA 25 212 76 18

## 6 NA 20 182 63 18

## Proc\_major\_therapeutic Proc\_major\_diagnostic Use\_any\_intens\_care Use\_ICU

## 1 22 NA 144 67

## 2 19 NA 137 61

## 3 20 NA 148 63

## 4 17 NA 96 47

## 5 17 NA 121 49

## 6 18 NA 95 38

## Use\_CCU Use\_NICU Use\_other\_care\_units Use\_mech\_ventilation

## 1 90 NA 90 23

## 2 88 NA 88 17

## 3 94 13 94 21

## 4 76 NA 76 19

## 5 86 NA 86 21

## 6 74 NA 74 16

## CIR009\_myocardial\_infar CIR017\_cardiac\_dysrhy CIR019\_heart\_failur

## 1 NA NA 16

## 2 NA NA 18

## 3 NA NA 21

## 4 NA NA 12

## 5 NA NA 14

## 6 NA NA 13

## INF002\_septicemia RSP002\_pneumonia RSP005\_acute\_bronchitis

## 1 42 28 61

## 2 38 29 50

## 3 66 36 70

## 4 36 21 46

## 5 34 14 35

## 6 29 20 21

## RSP008\_chronic\_pulmonary RSP009\_asthma RSP010\_aspiration\_pneumonitis

## 1 81 NA NA

## 2 73 NA NA

## 3 83 NA NA

## 4 51 NA NA

## 5 65 NA NA

## 6 54 NA NA

## RSP012\_respiratory\_failure All\_other\_conditions Missing\_pay Hosp\_L\_metro

## 1 32 77 NA NA

## 2 39 93 NA NA

## 3 31 89 NA NA

## 4 40 65 NA NA

## 5 51 54 NA NA

## 6 30 63 NA NA

## Race\_not\_in\_SID area X Y geom YEAR REGION DIVISION STATE

## 1 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## 2 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## 3 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## 4 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## 5 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## 6 NA NA NA NA MULTIPOLYGON EMPTY 2017 4 9 2

## POPULATION

## 1 740983

## 2 740983

## 3 740983

## 4 740983

## 5 740983

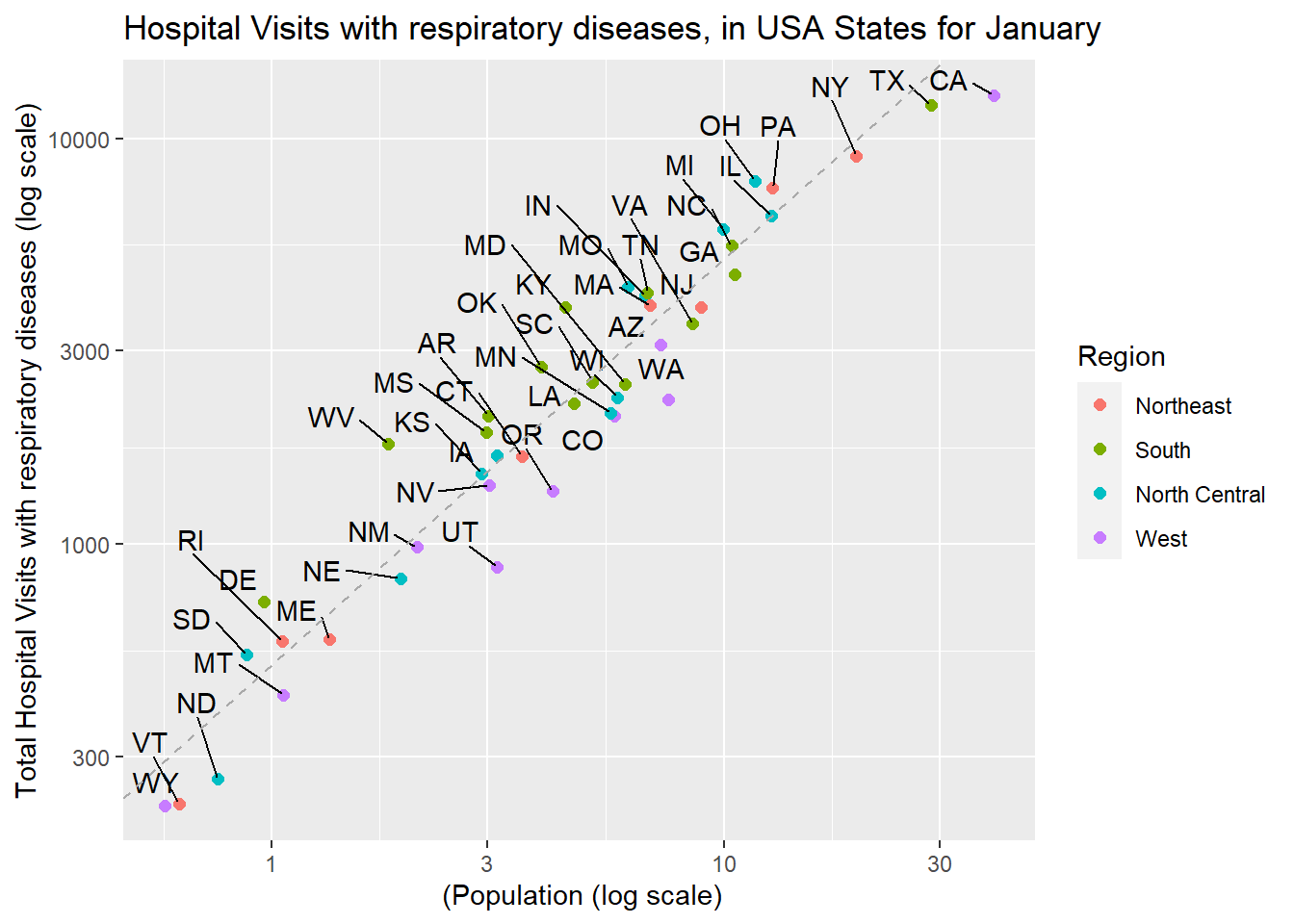
## 6 740983

To prepare the merging of DataSets 1 and 2 we should first aggregate AQI data by US State and month, having media from: Cat, AQI

### Visual analysis for Hospital visits with respiratory diseases aggretated Data by Month

For January Data

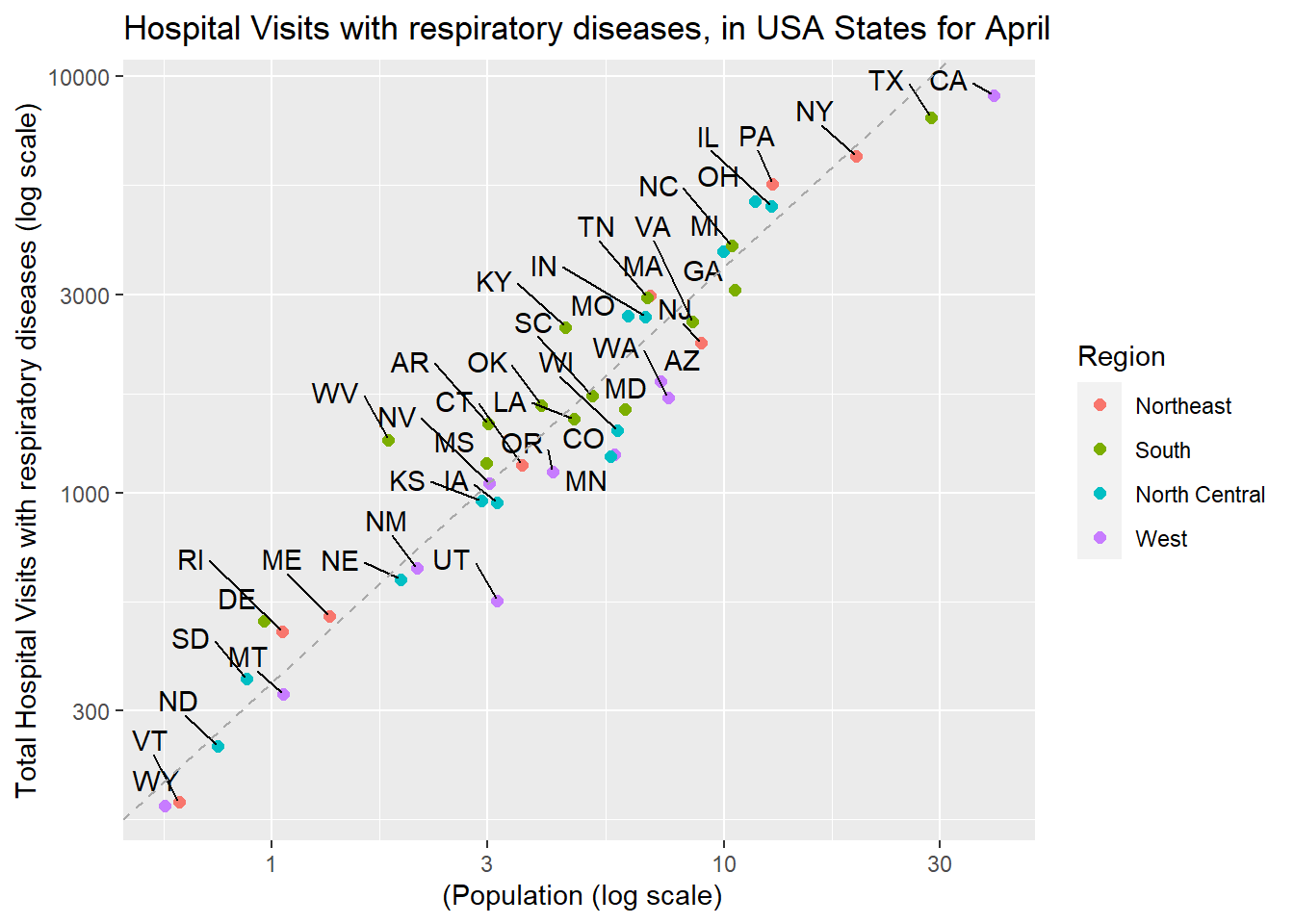
Visits\_g1



### Visual analysis for Hospital visits with respiratory diseases aggretated Data by Month

For April Data

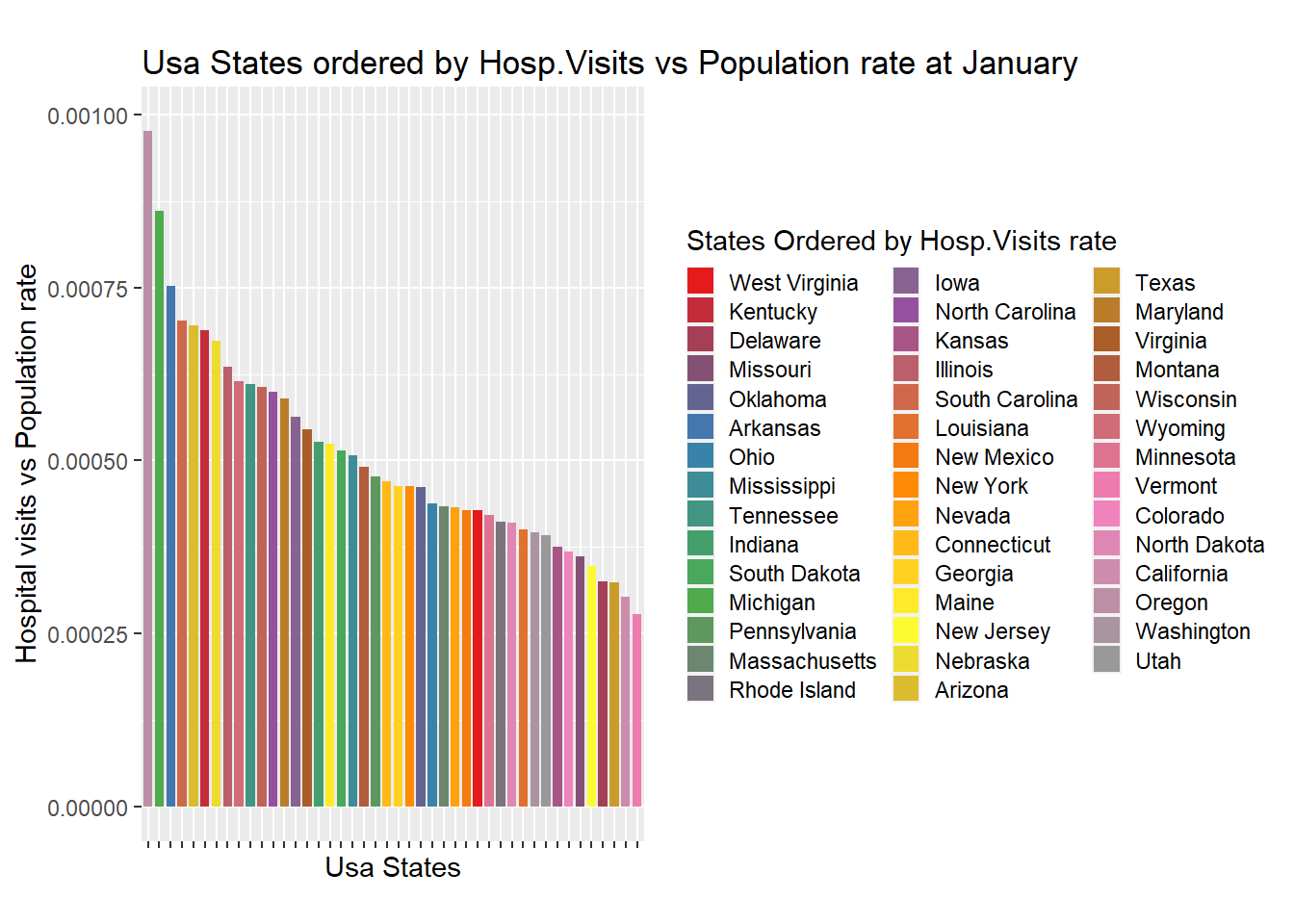
Visits\_g4



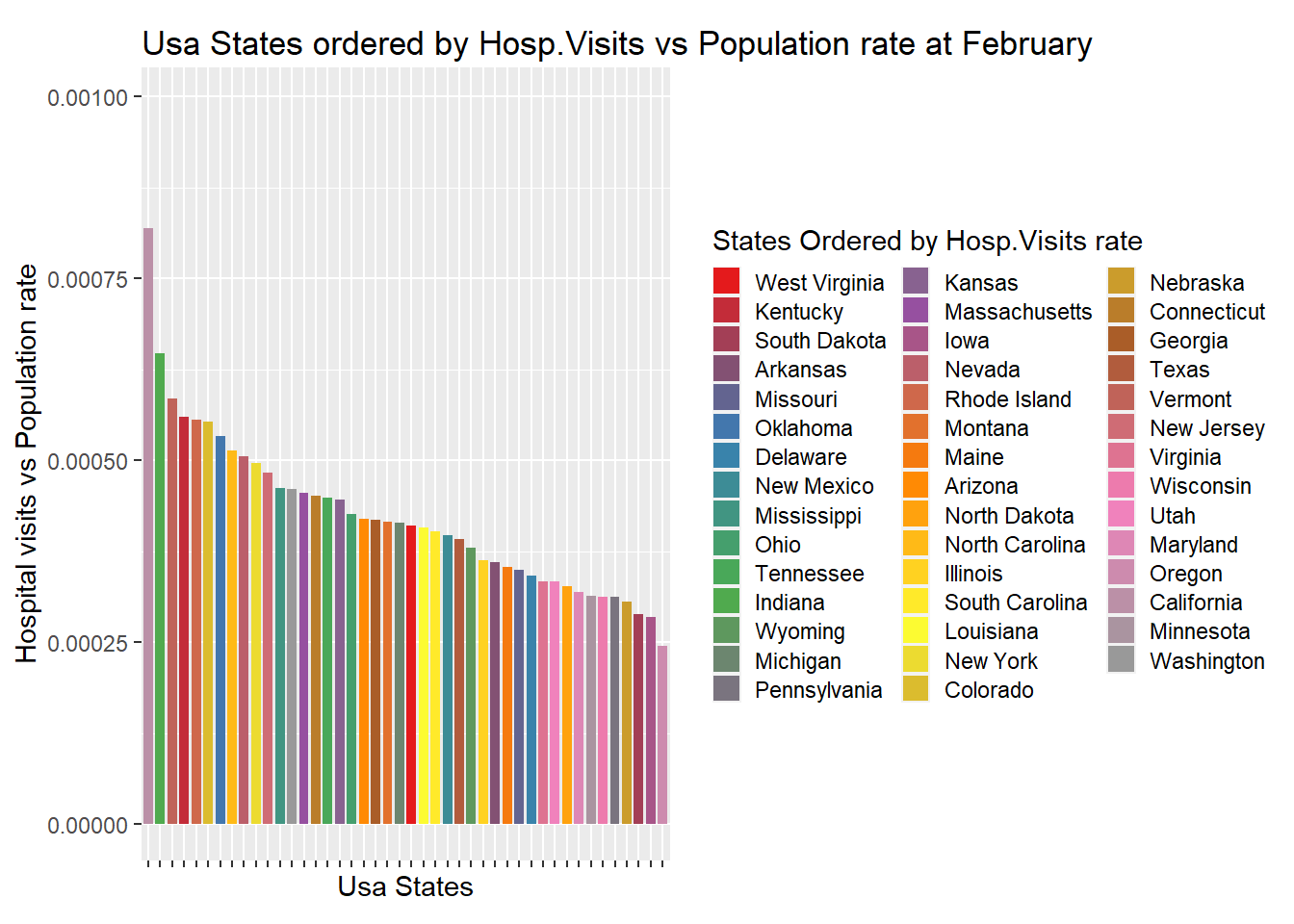
We can see that in general there are a few US State far from median line

### Now plotting states and Hosp.Visit rates by Month

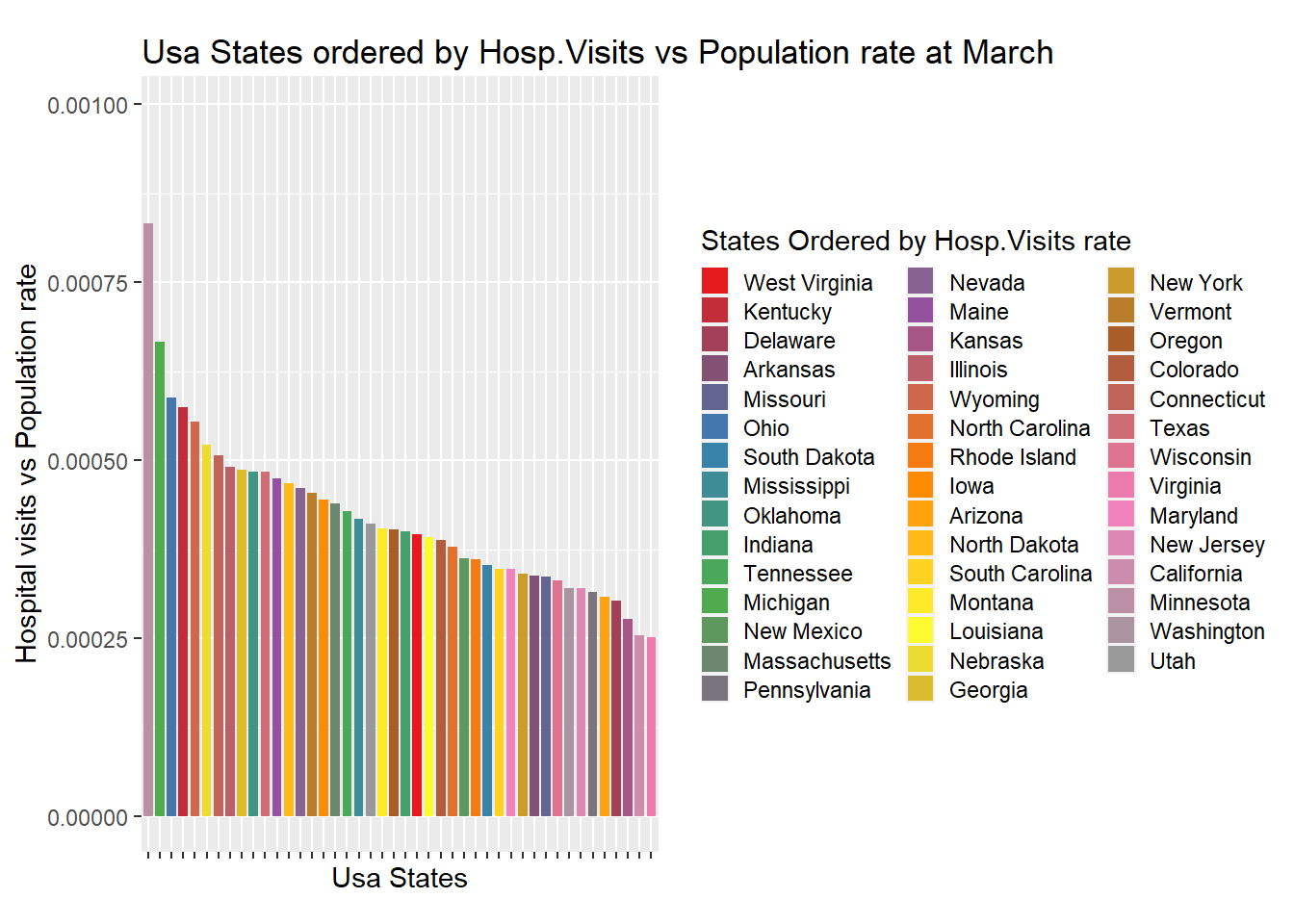
states\_graph\_1



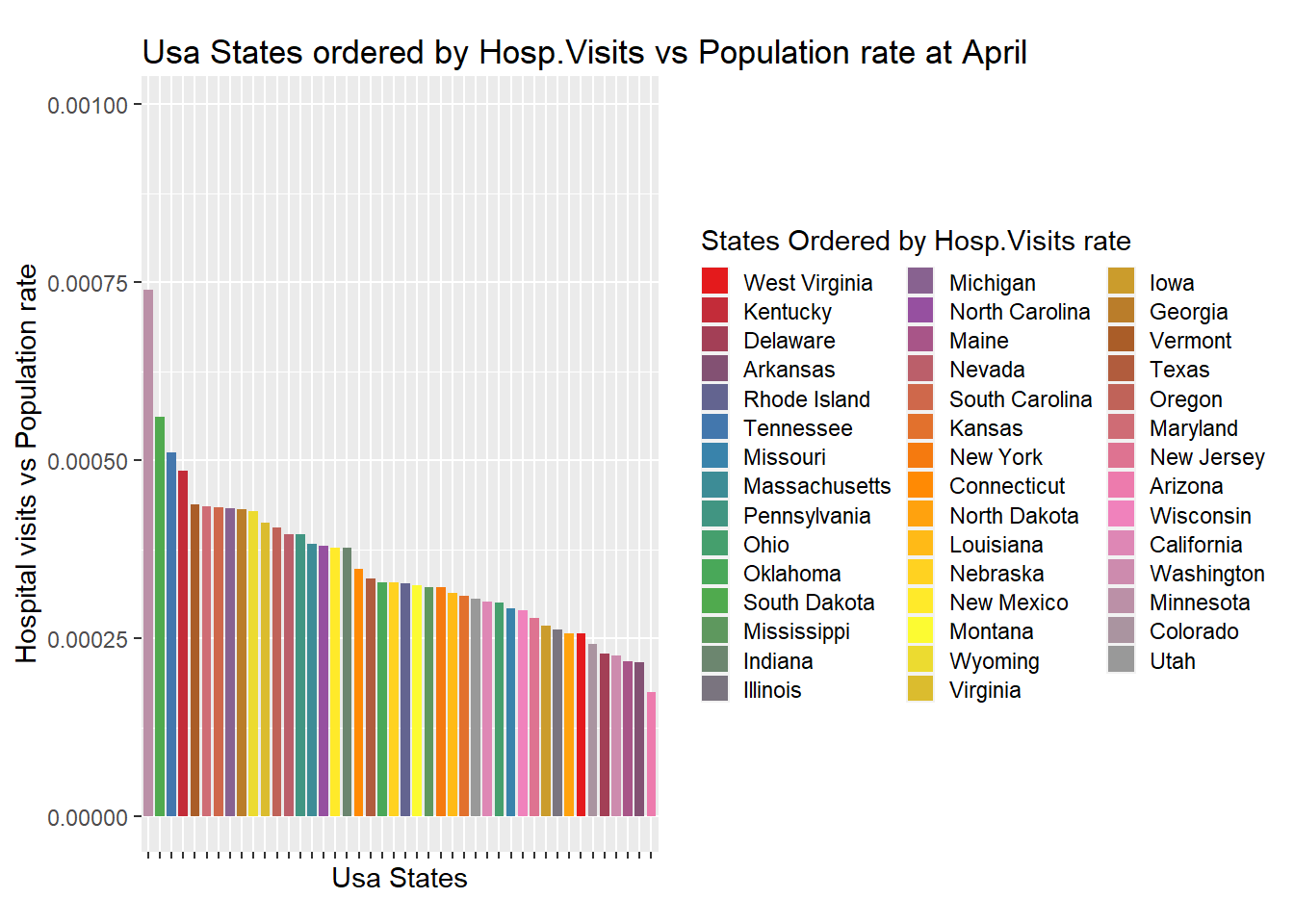
states\_graph\_2



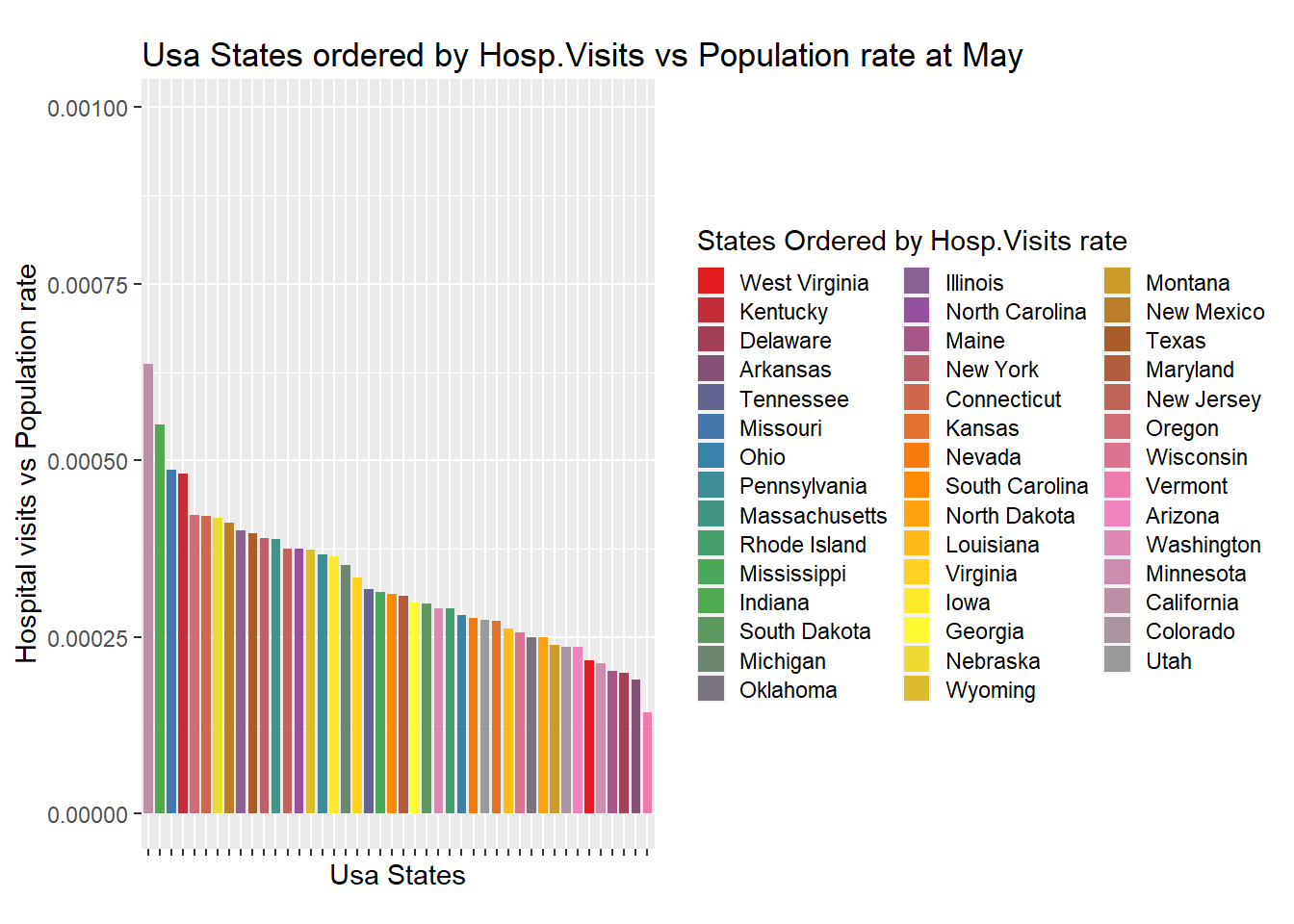
states\_graph\_3



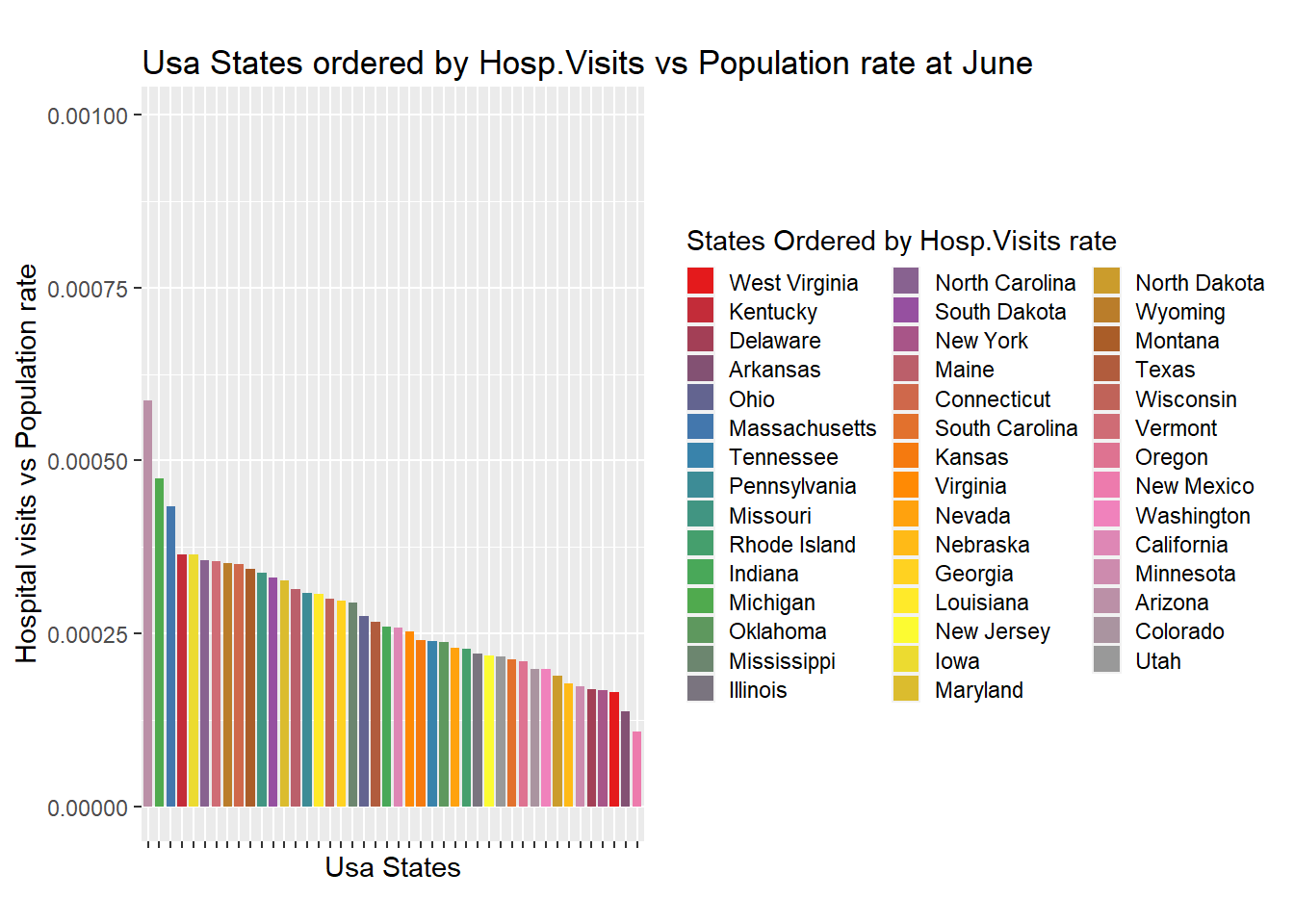
states\_graph\_4



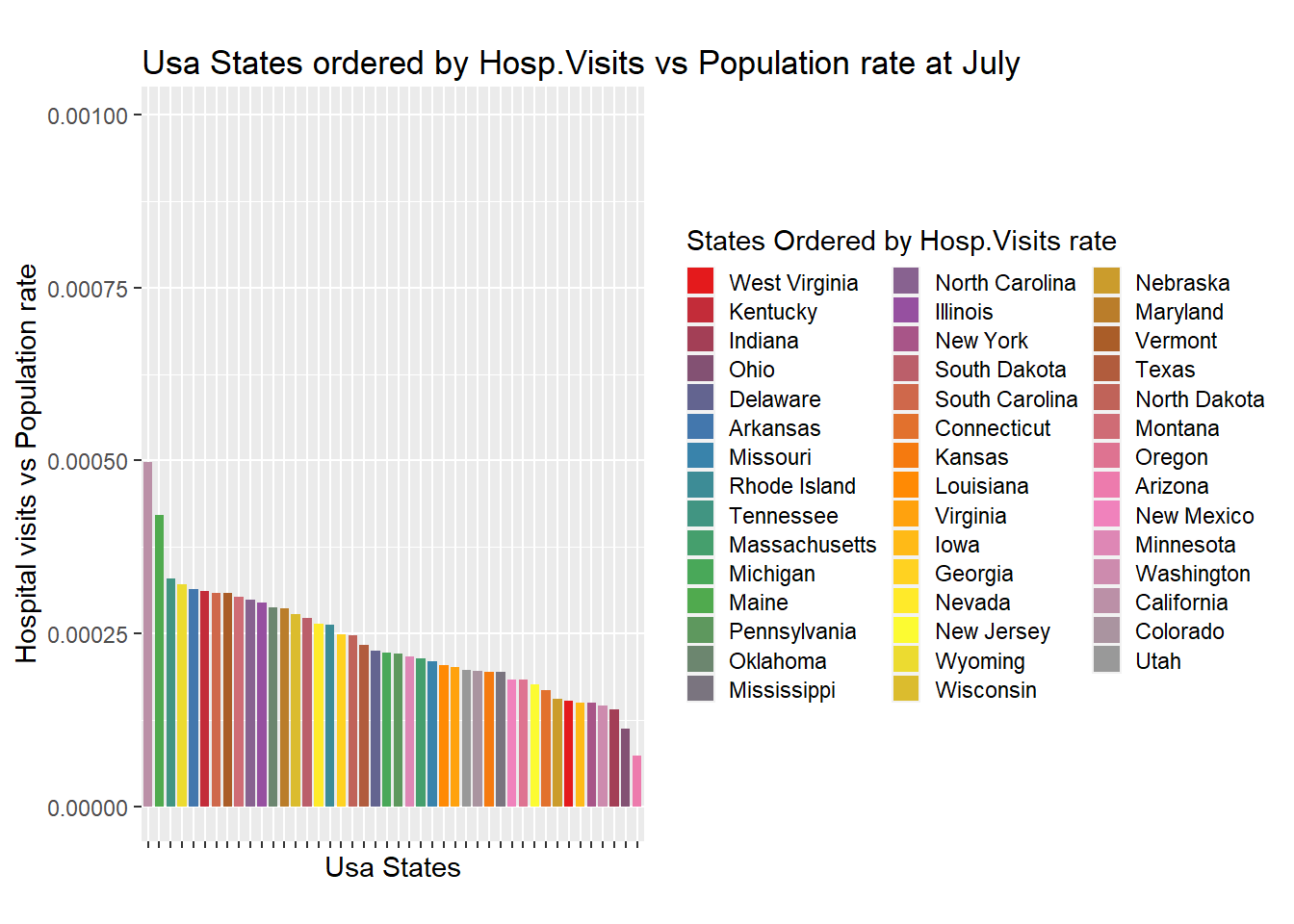
states\_graph\_5



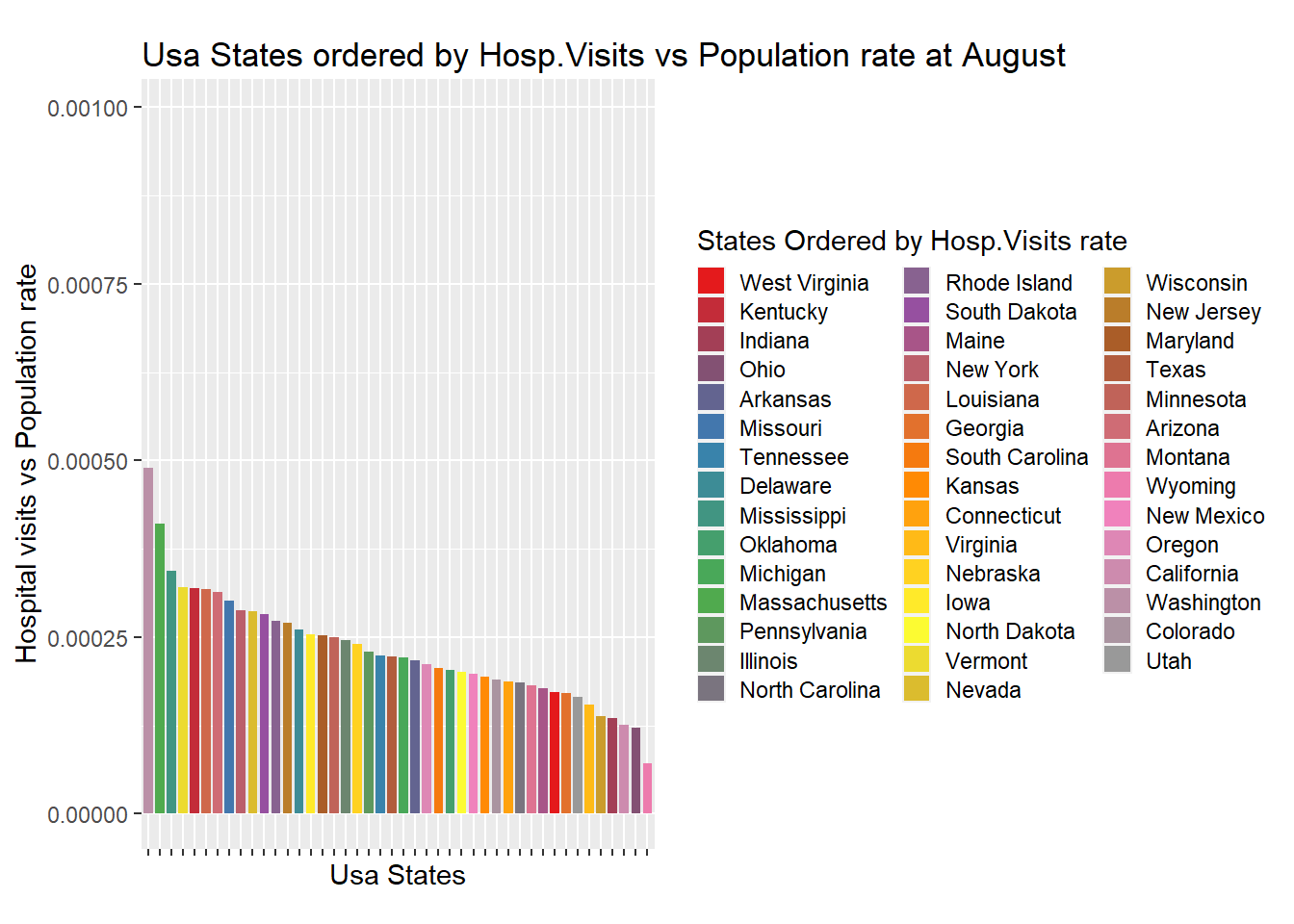
states\_graph\_6



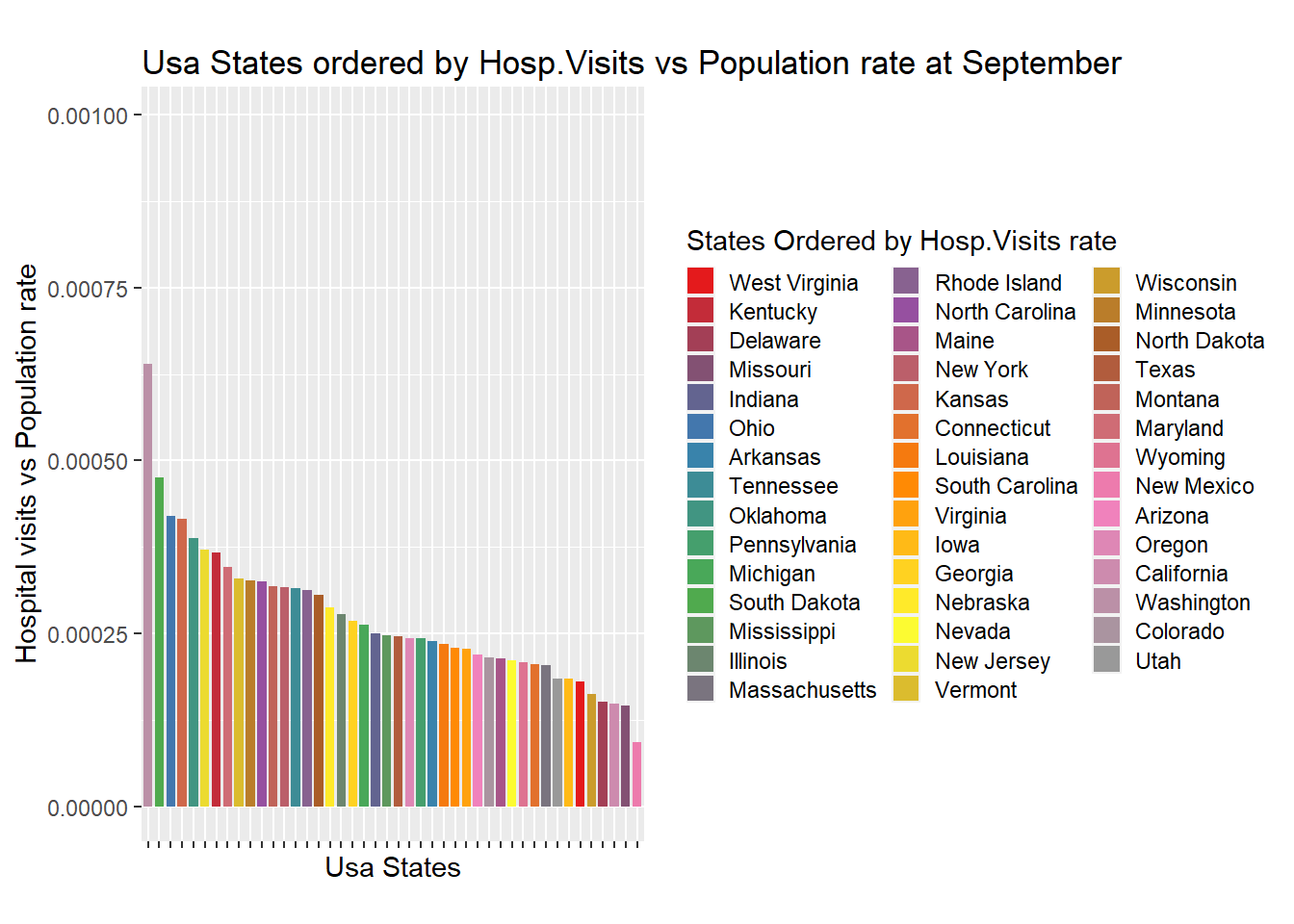
states\_graph\_7



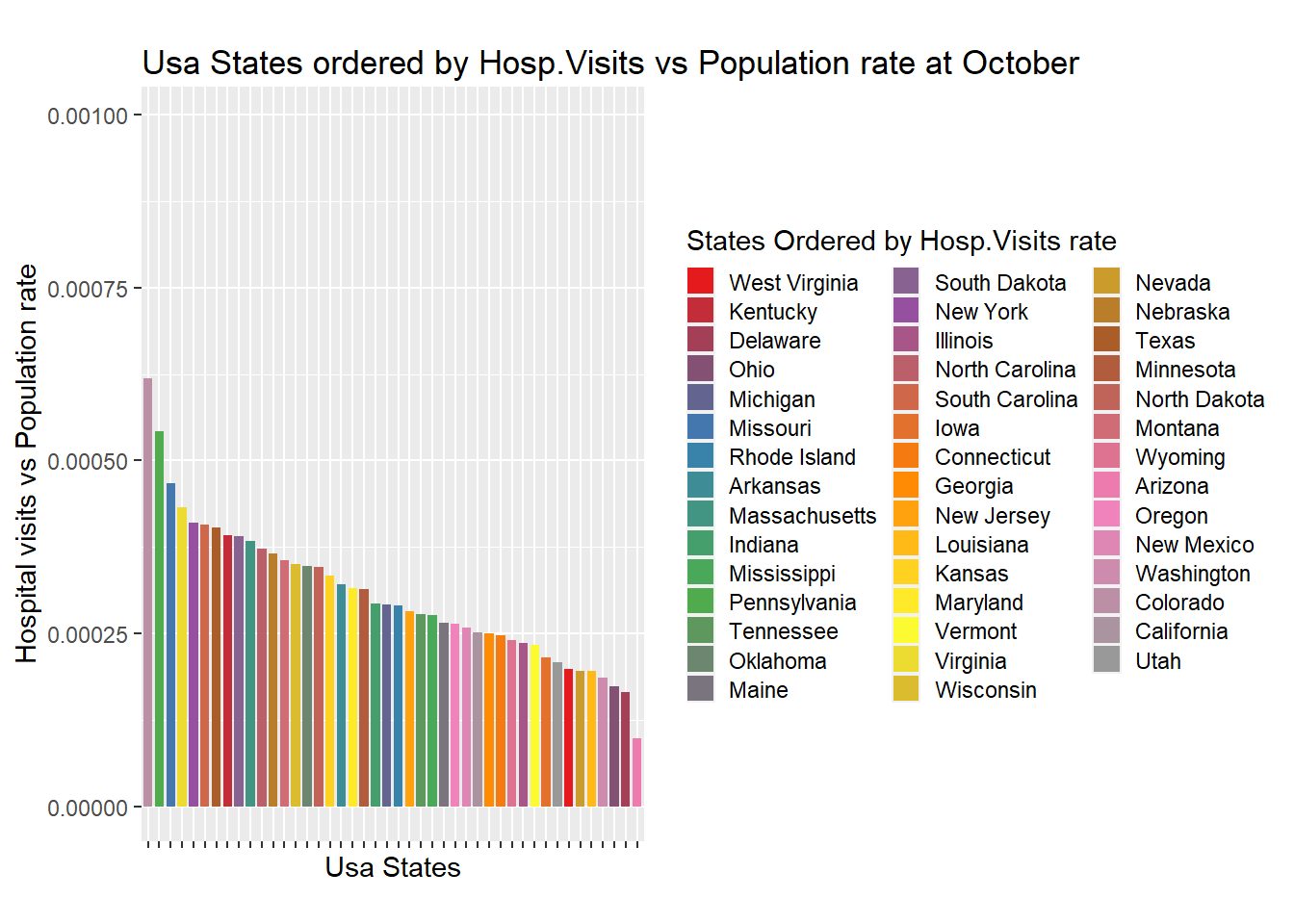
states\_graph\_8



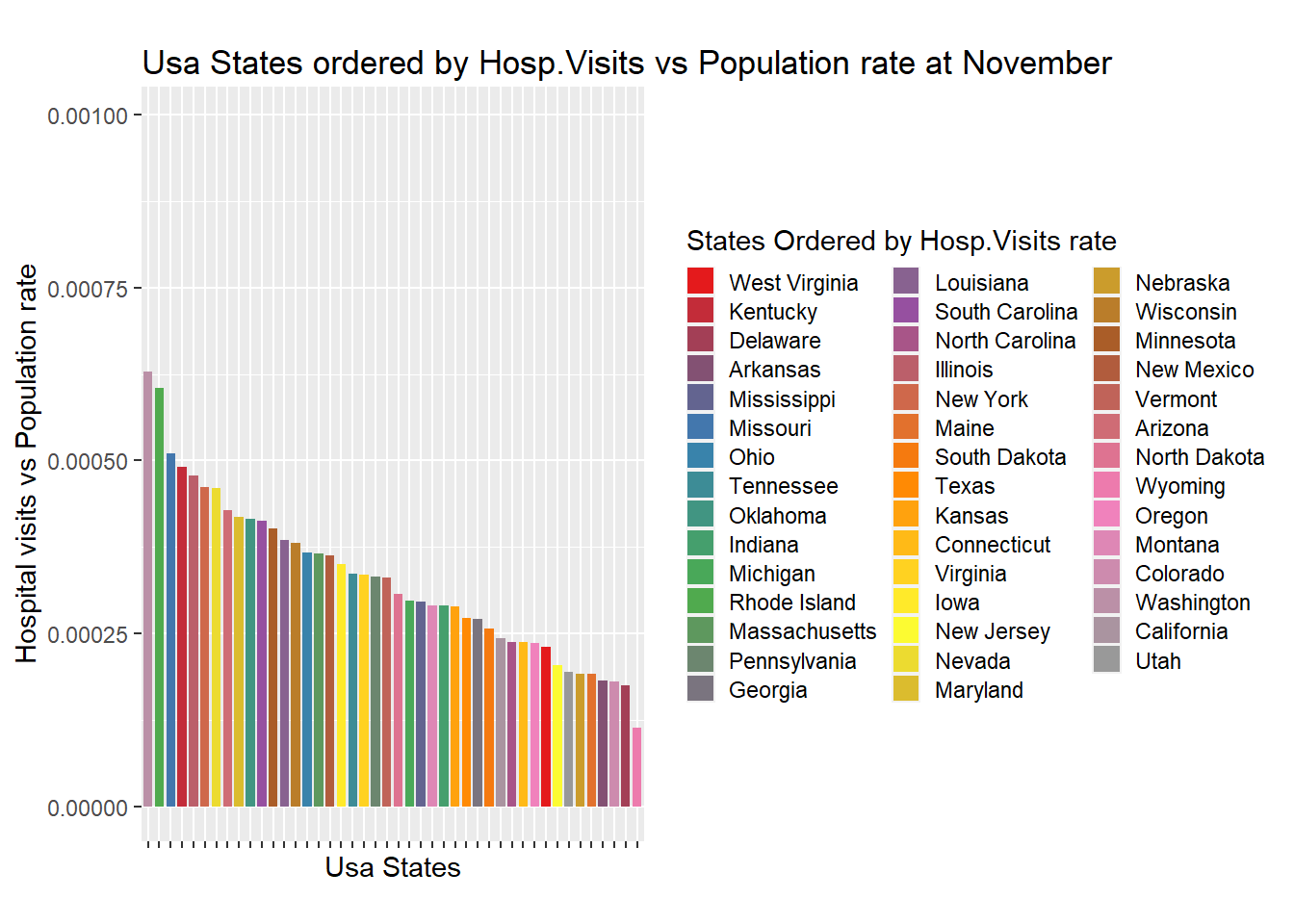
states\_graph\_9



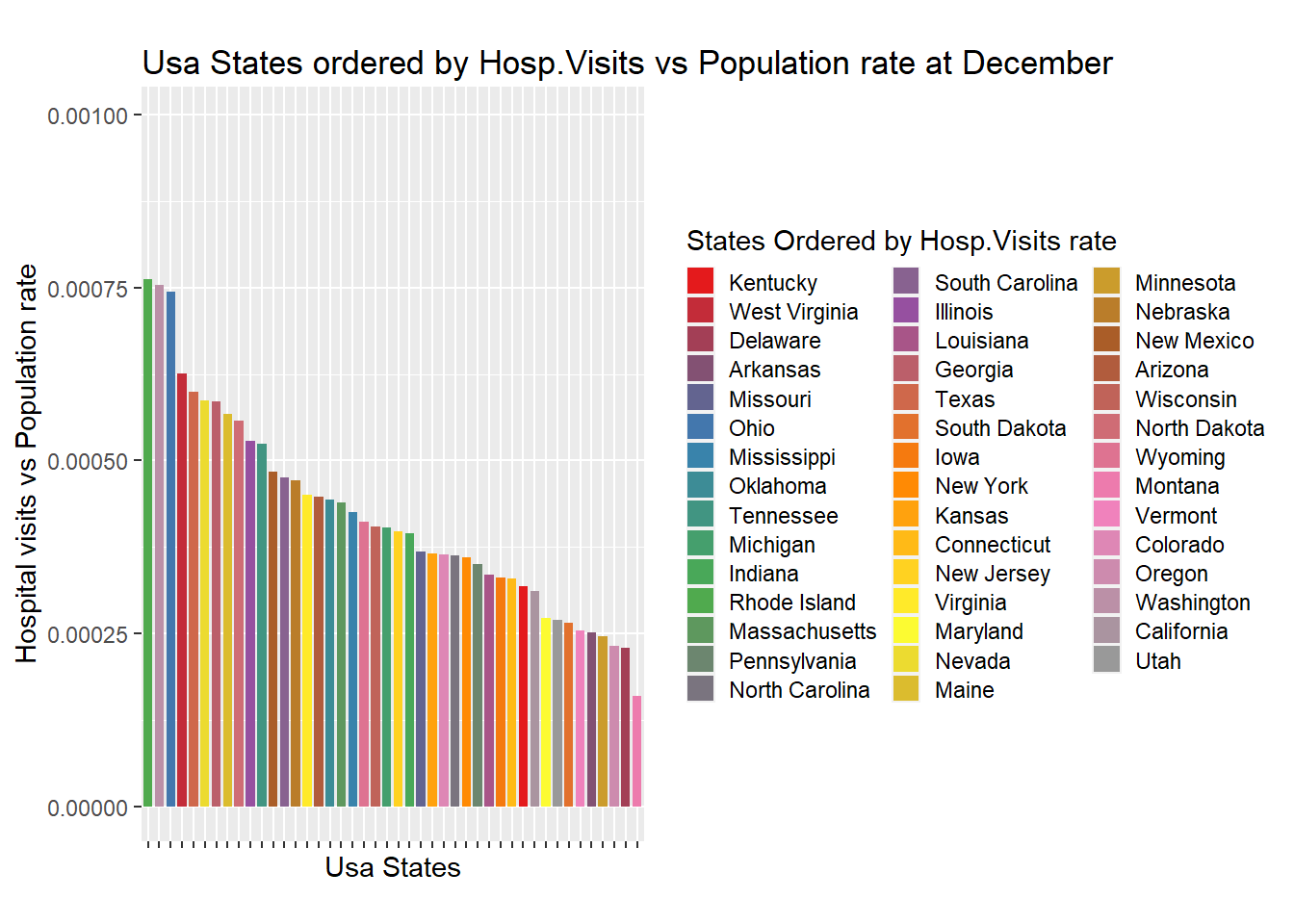
states\_graph\_10



states\_graph\_11



states\_graph\_12



We can see that 10 Sates always repeat with high rates of Hosp.Visits each year from 2017 to 2020

### Analyzing Air Quality Data on USA States

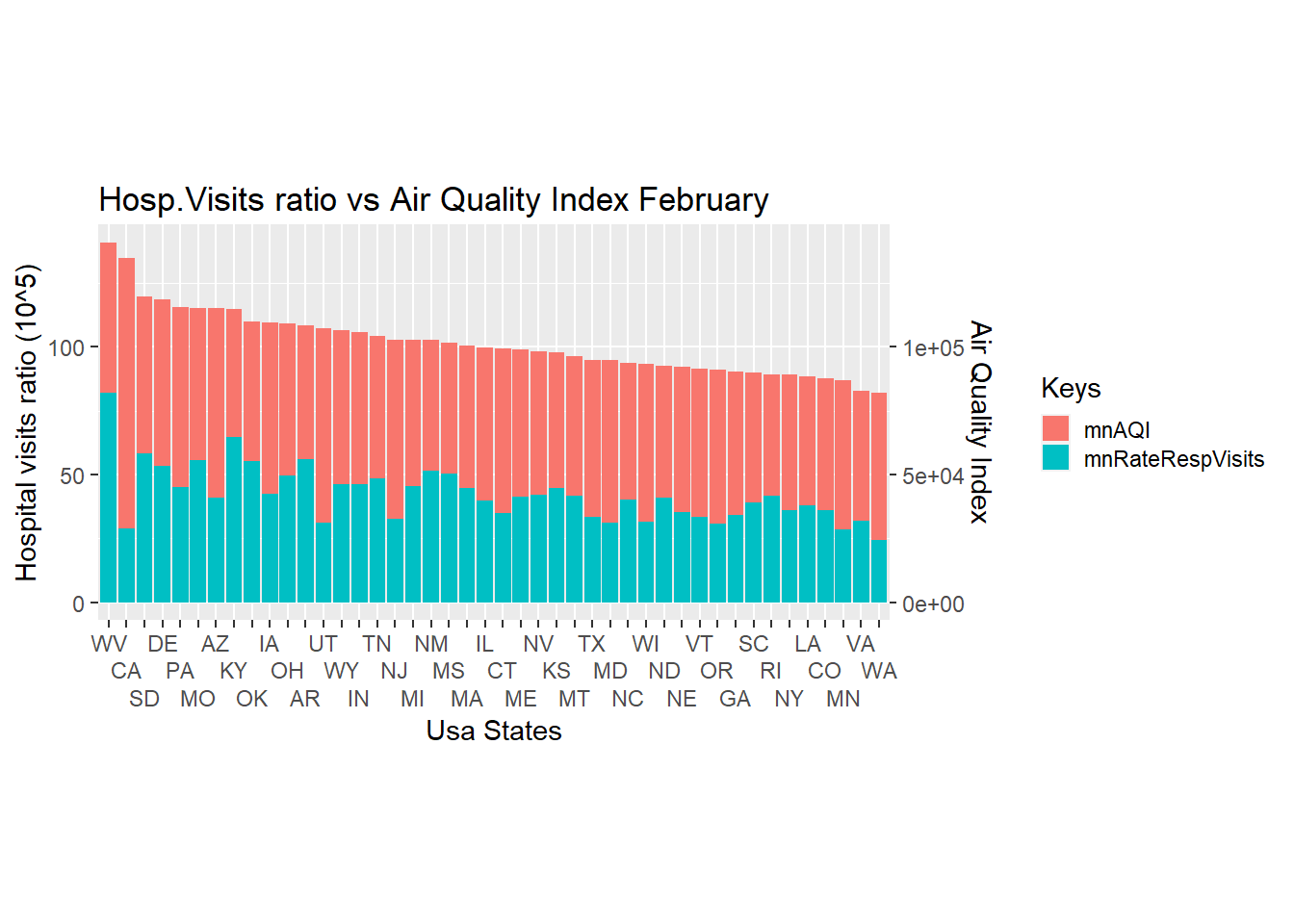
Standard deviation and mean rate of Air Quality Indexes per month and USA States

### Presenting Hospital Visits rate vs Air Quality Indexes in all months

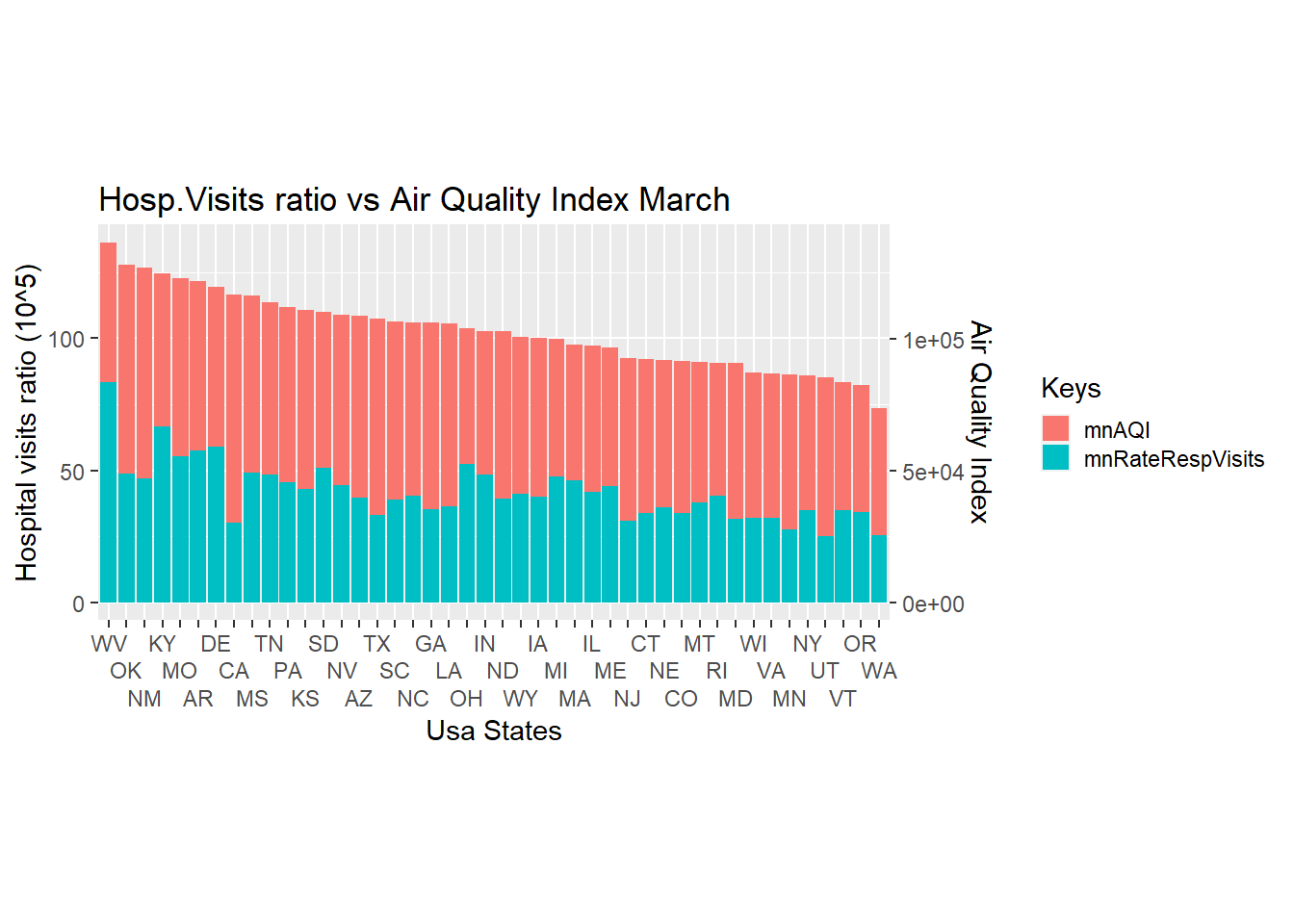
states\_tot\_1



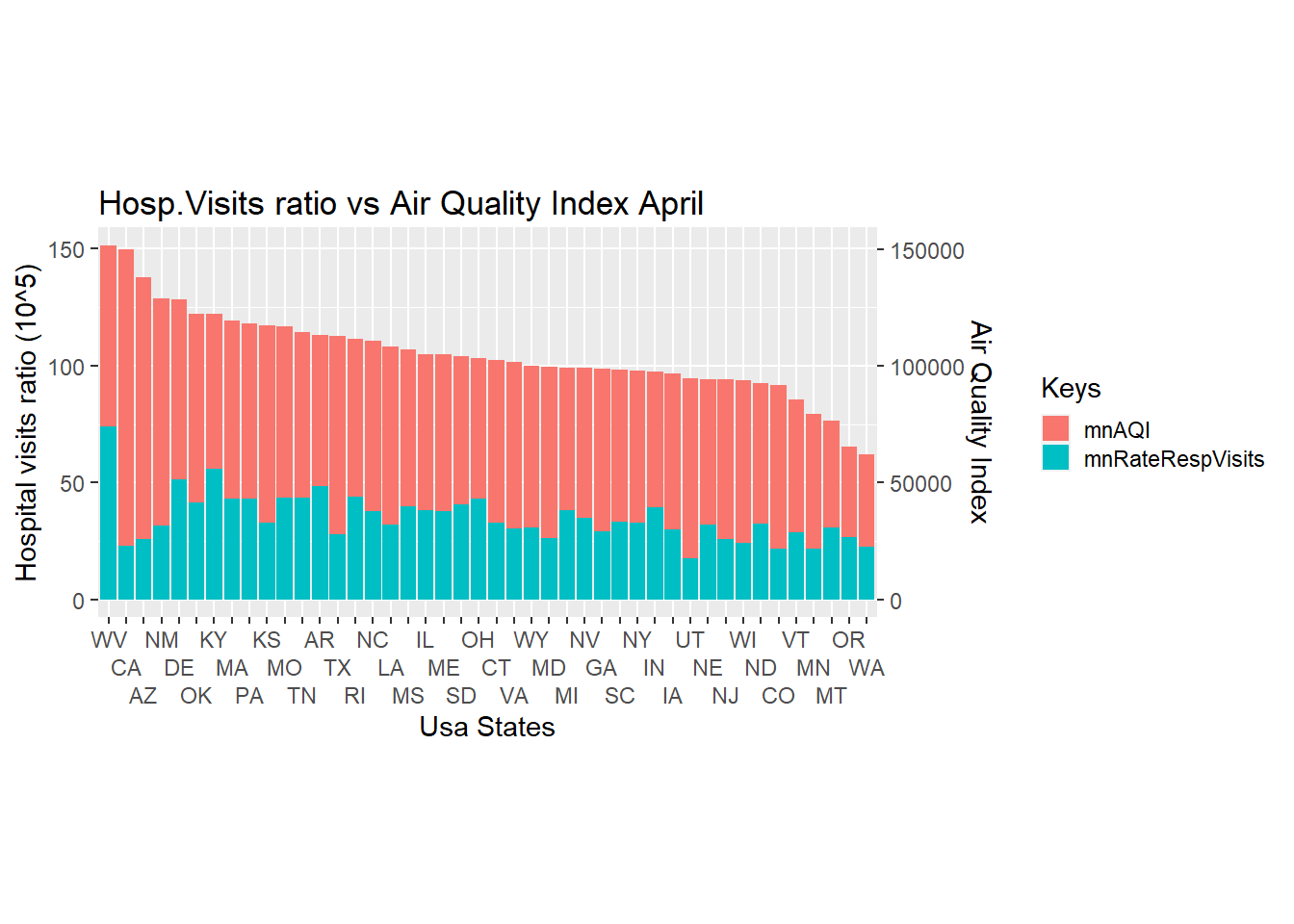
states\_tot\_2



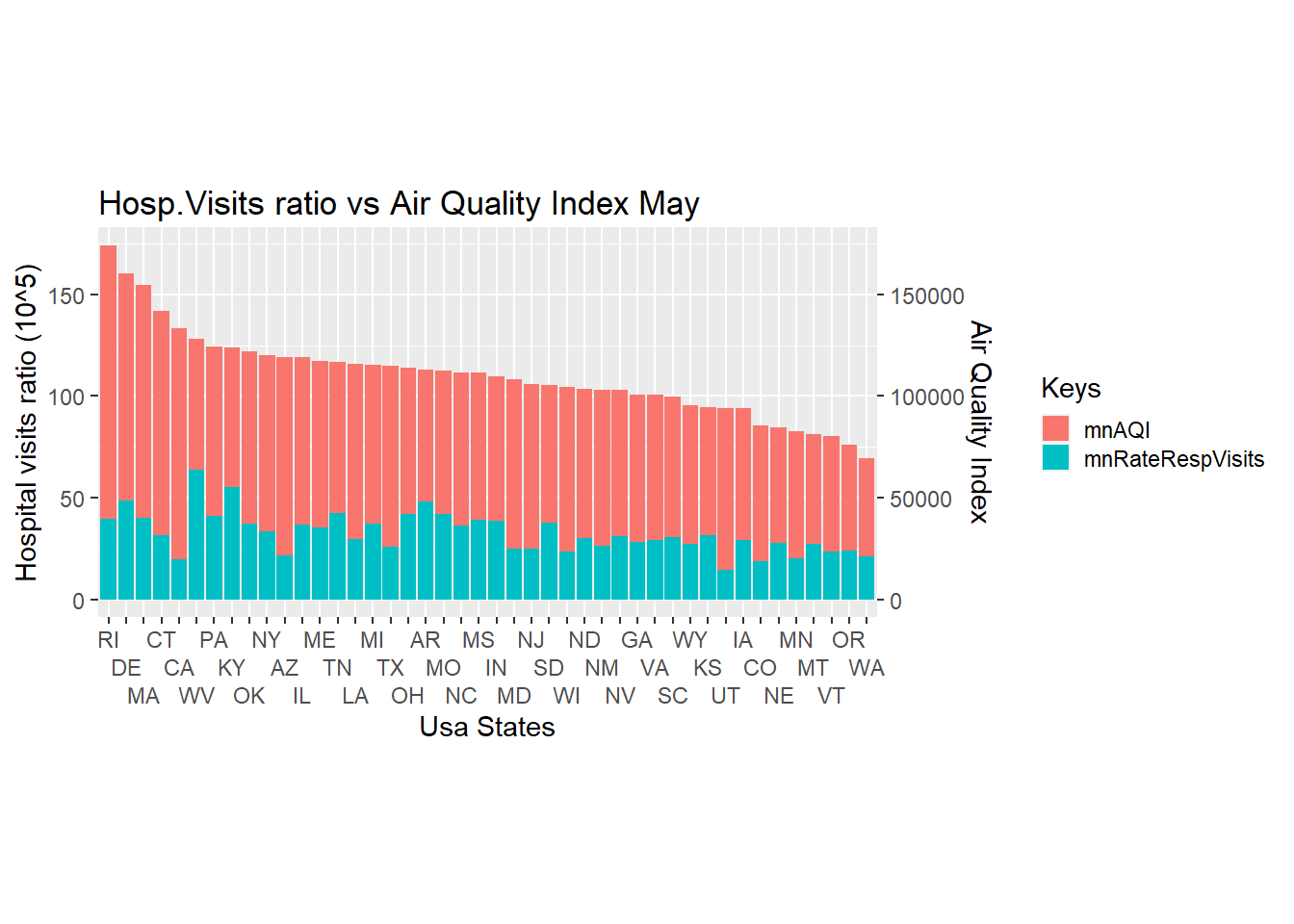
states\_tot\_3



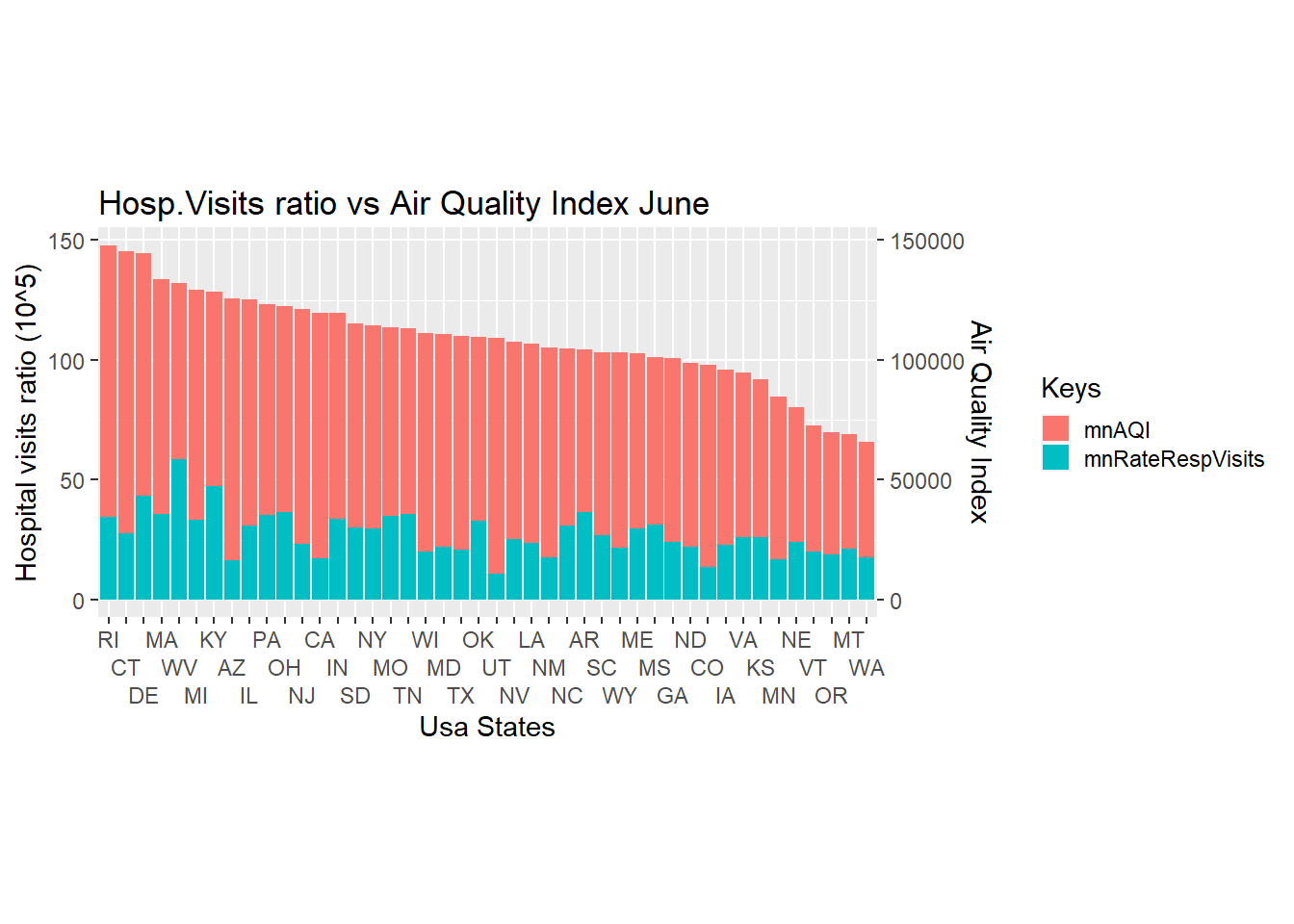
states\_tot\_4



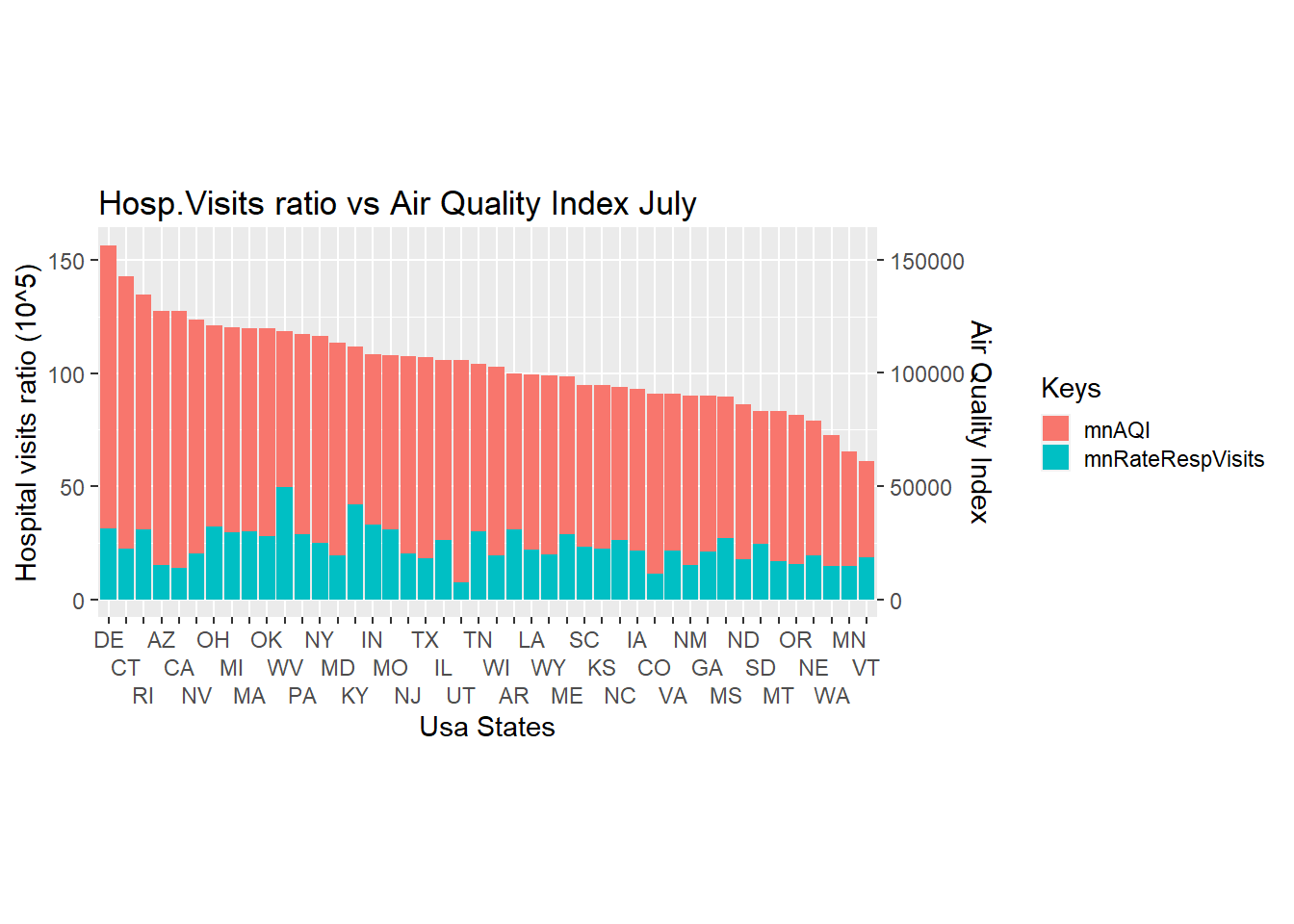
states\_tot\_5



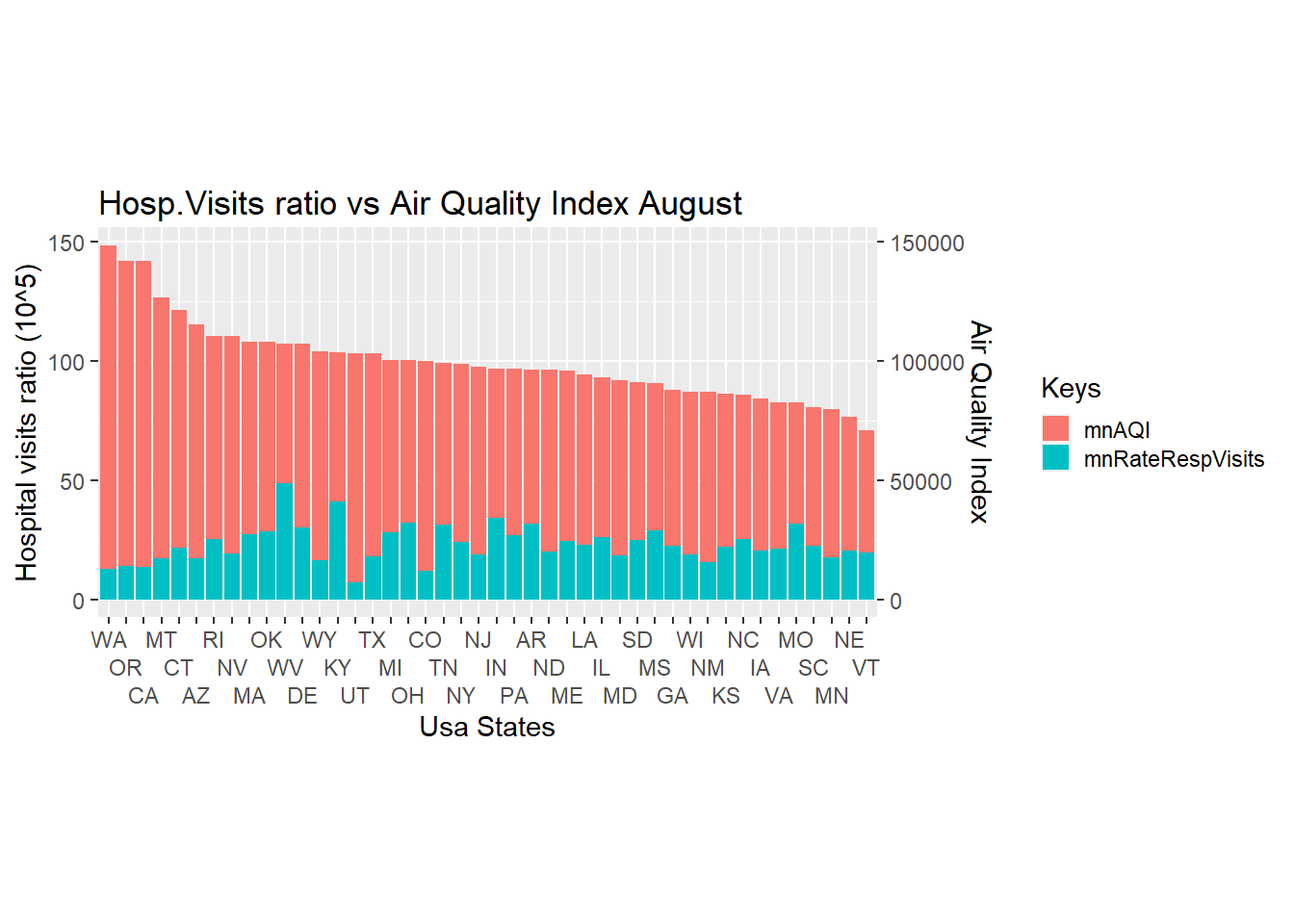
states\_tot\_6



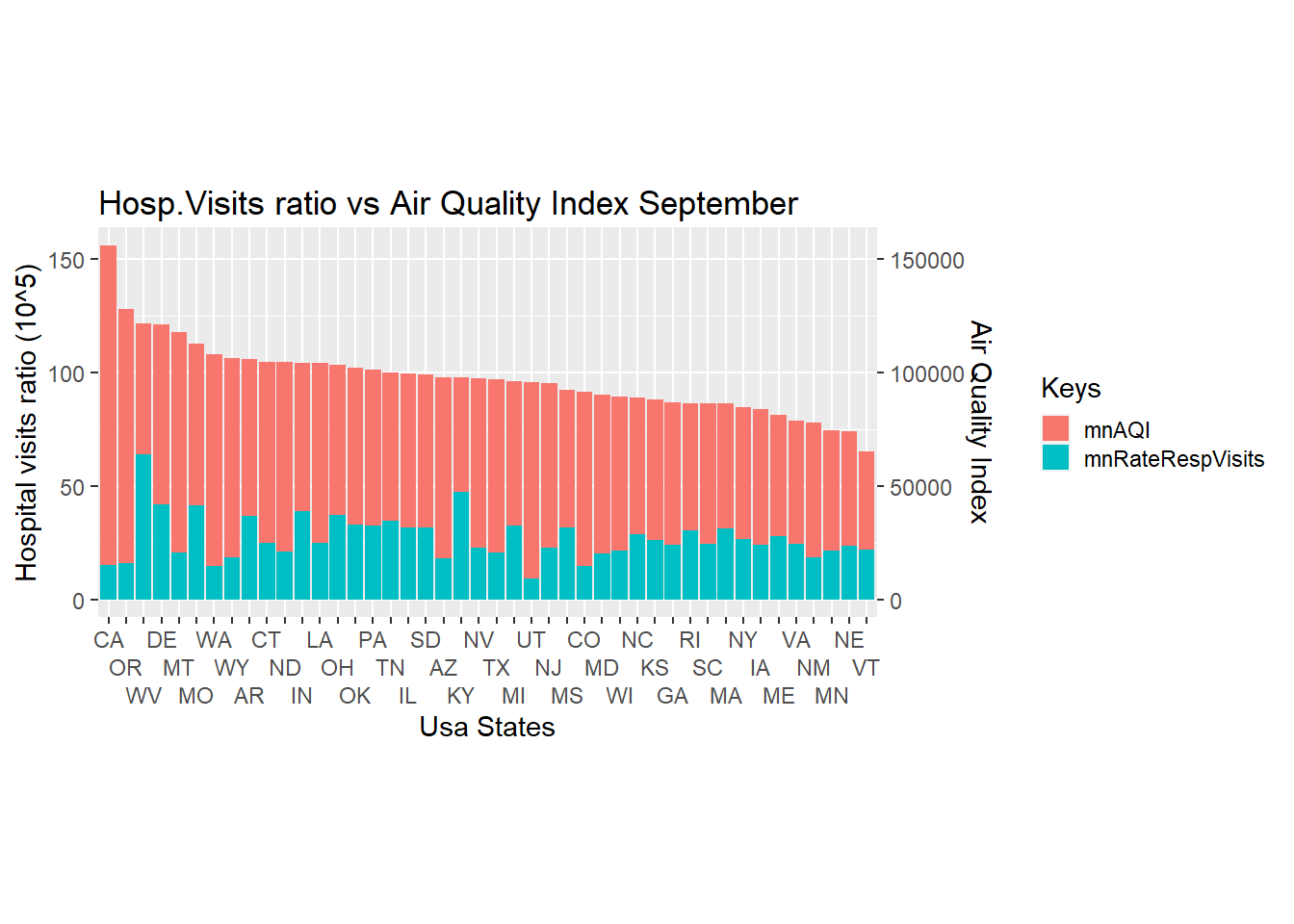
states\_tot\_7



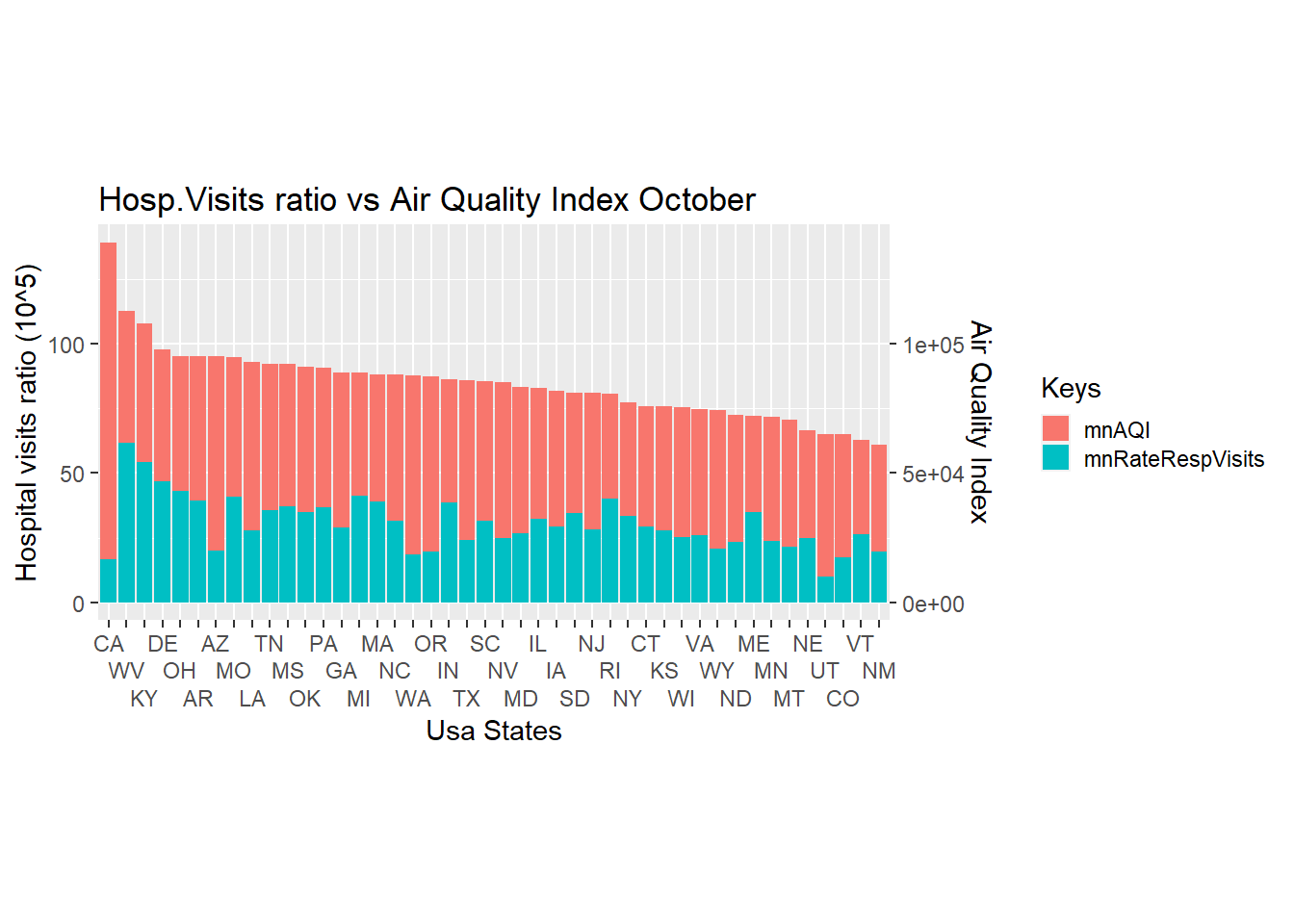
states\_tot\_8



states\_tot\_9



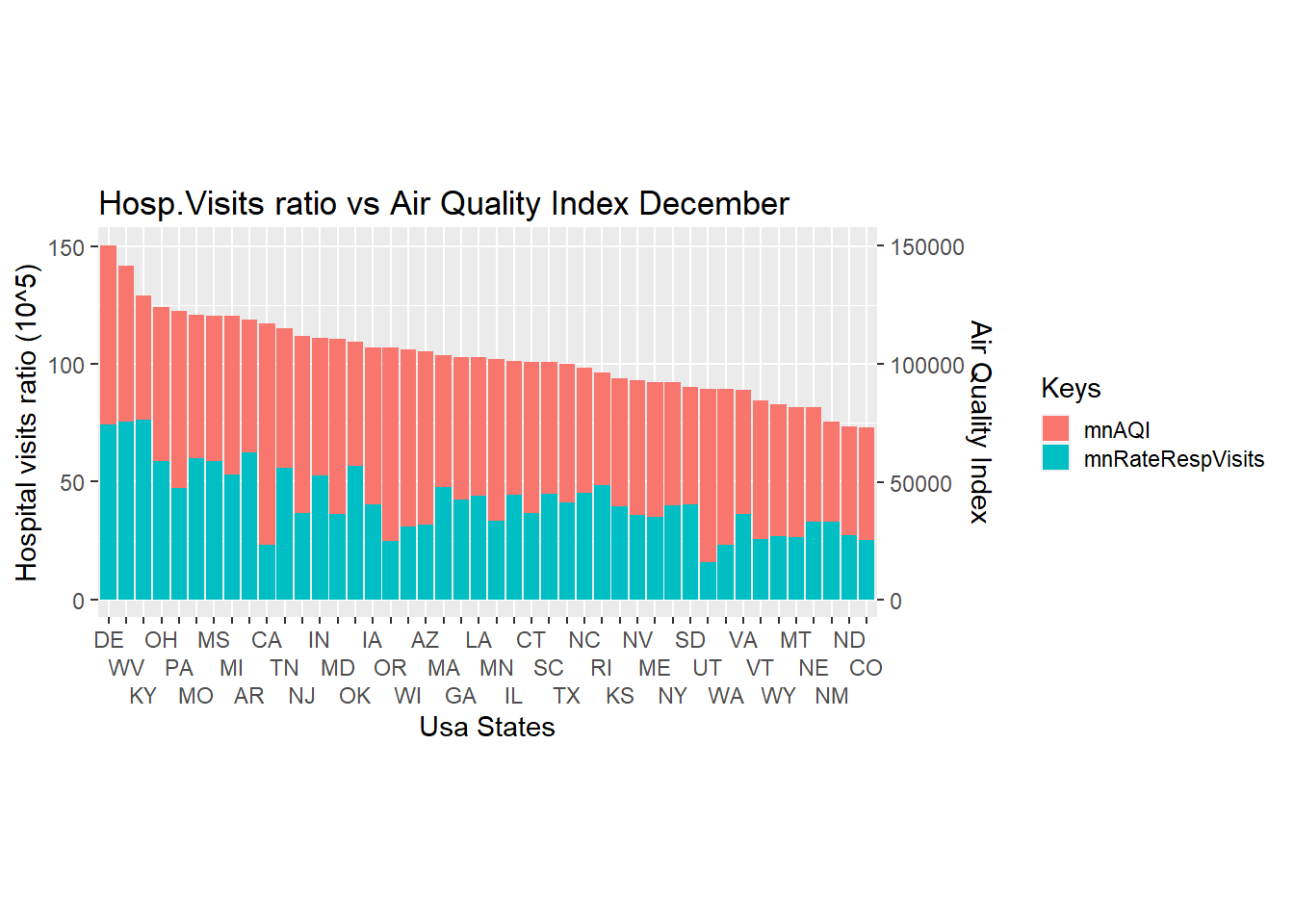
states\_tot\_10



states\_tot\_11



states\_tot\_12



Its is clear than there is a correlation between Hospital Visits ratio and the Air Quality Index every month of the year Saving Aggregated Data for the next Notebook

# Saving in a data file the Tibble with Hospital data Visits and States population:

save(TotData, file = paste(datapath,file\_TotData\_save,sep=""))

# Features selecction

#### Marcos Mariscal

#### 23/5/2021

#### This RMarkdown tries to find correlations between features to select the best group for the goal of this project

### Loading Data field of Hosp.Visits and AQI data

#### Replace Zeros to NA for all numeric columns

#### Complete NA values, we will use predictive mean matching (pmm) (multi nominal logistic regression)

### Distribution of imputed values with observed ones for Hosp.visits per age ranges

p1 <- ggplot(Completed) + geom\_point(aes(statabb,Ages\_0\_4))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "USA States")+

ggtitle("Hosp.Visits in Ages ranges")

p2 <- ggplot(Completed) + geom\_point(aes(statabb,Ages\_5\_9))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "USA States")+

ggtitle("Hosp.Visits in ages ranges")

p3 <- ggplot(Completed) + geom\_point(aes(statabb,Ages\_10\_17))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "USA States")+

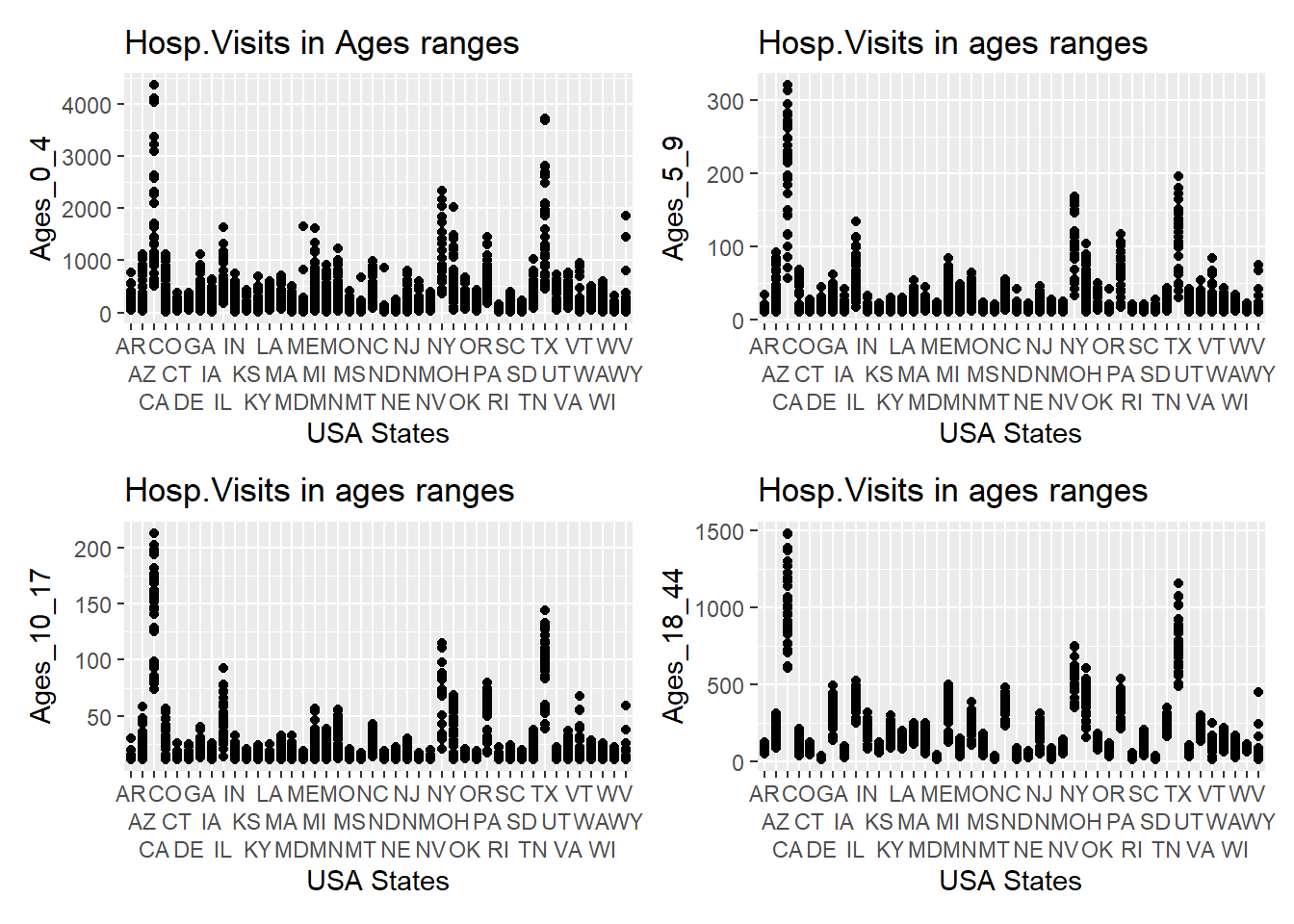
ggtitle("Hosp.Visits in ages ranges")

p4 <- ggplot(Completed) + geom\_point(aes(statabb,Ages\_18\_44))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "USA States")+

ggtitle("Hosp.Visits in ages ranges")

p1 + p2 + p3 + p4



### Distribution of imputed values with observed ones for different respiratori ilnesses

p5 <- ggplot(Completed) + geom\_point(aes(MONTH,RSP002\_pneumonia))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "Months")+

ggtitle("Hospital Visits")

p6 <- ggplot(Completed) + geom\_point(aes(MONTH,RSP005\_acute\_bronchitis))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "Months")+

ggtitle("Hospital Visits")

p7 <- ggplot(Completed) + geom\_point(aes(MONTH,RSP009\_asthma))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "Months")+

ggtitle("Hospital Visits")

p8 <- ggplot(Completed) + geom\_point(aes(MONTH,RSP010\_aspiration\_pneumonitis))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "Months")+

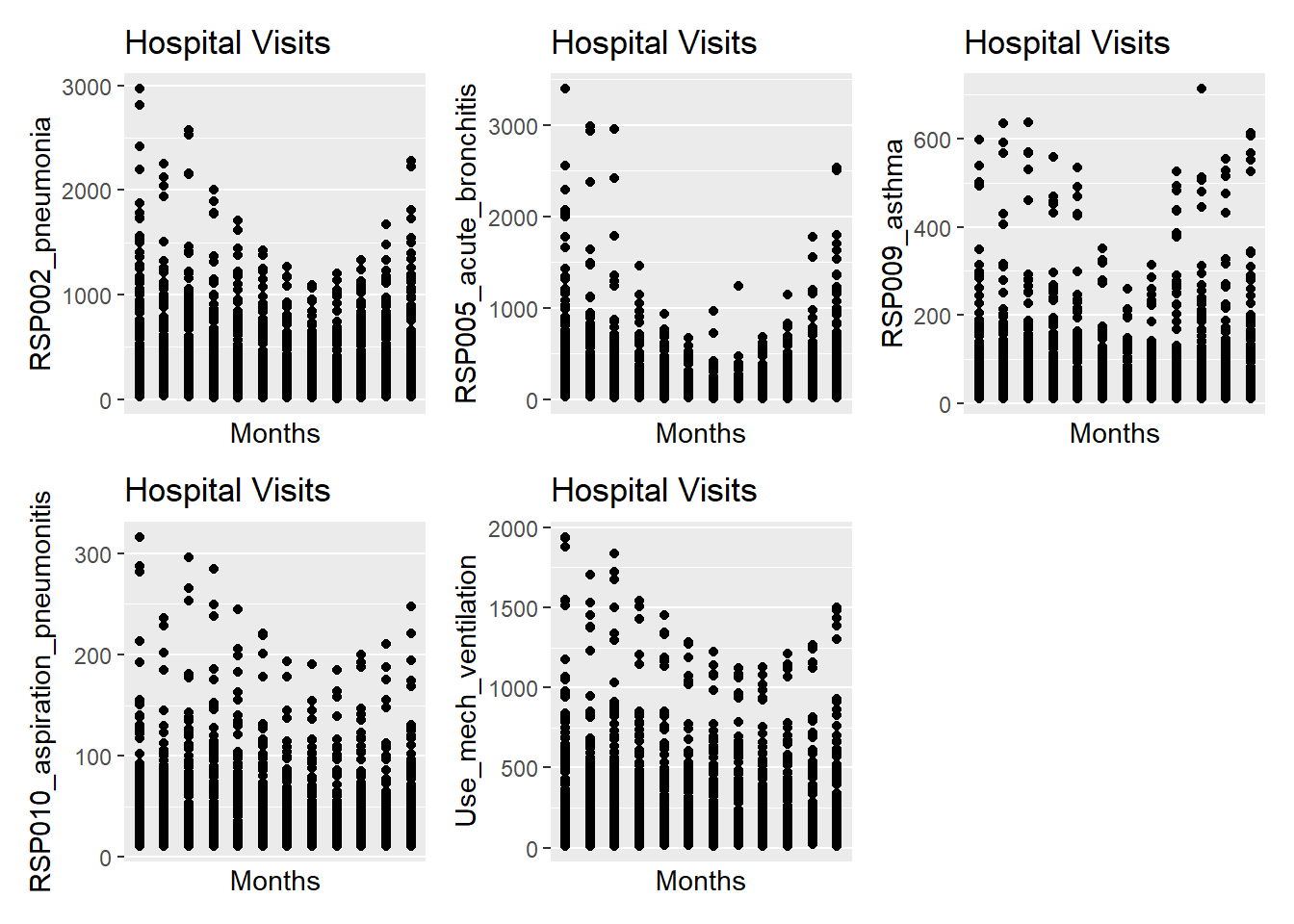
ggtitle("Hospital Visits")

p9 <- ggplot(Completed) + geom\_point(aes(MONTH,Use\_mech\_ventilation))+

scale\_x\_discrete(guide = guide\_axis(n.dodge=3), name = "Months")+

ggtitle("Hospital Visits")

p5 + p6 + p7 + p8 + p9



AggTotT1 <- TotRespData%>%select (-c(MONTH,Ages\_0\_4, Ages\_5\_9, Ages\_10\_17, Ages\_18\_44, RSP005\_acute\_bronchitis, RSP002\_pneumonia, RSP009\_asthma, RSP010\_aspiration\_pneumonitis,Use\_mech\_ventilation))

TotRespData <- inner\_join(x = AggTotT1, y =Completed , by = c("statabb","Dates"))

TotRespData<- TotRespData%>%select(STATES,statabb,Dates,MONTH,statreg,Category,Cat,AQI,Ages\_0\_4, Ages\_5\_9, Ages\_10\_17, Ages\_18\_44, Ages\_45\_64,Ages\_65\_79, Ages\_80, RSP002\_pneumonia, RSP005\_acute\_bronchitis, RSP008\_chronic\_pulmonary, RSP009\_asthma, RSP010\_aspiration\_pneumonitis, RSP012\_respiratory\_failure, Use\_mech\_ventilation, RateRespVisits, TotResp, POPULATION, All\_discharges)

### Correlation Matrix

#Correlation for Air Quality Index and all Hospital Visits of Respiratory illness

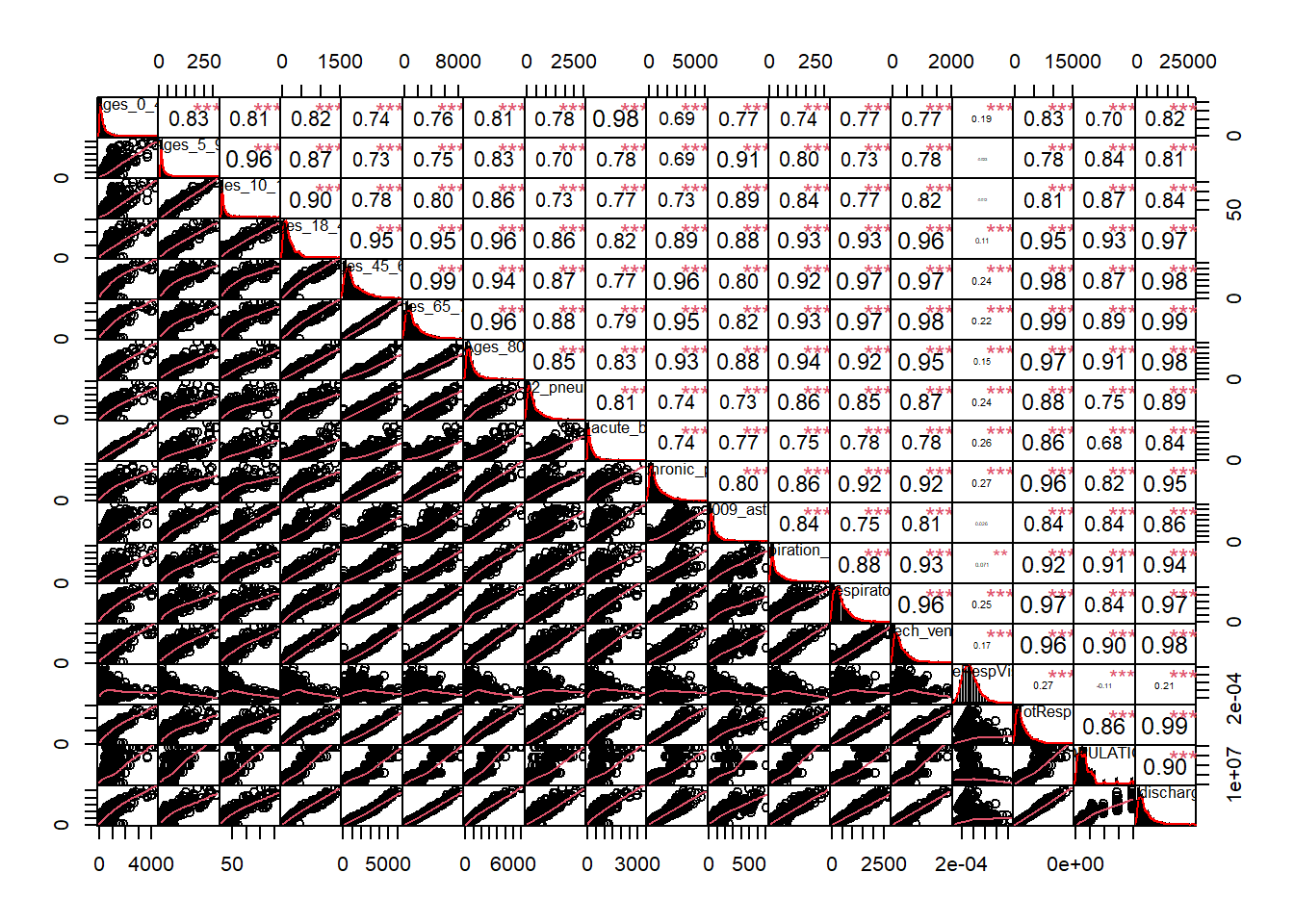
# ensure the results are repeatable

set.seed(7)

# calculate correlation aggtotd

correlationMatrix <- cor(TotRespData[,8:26])

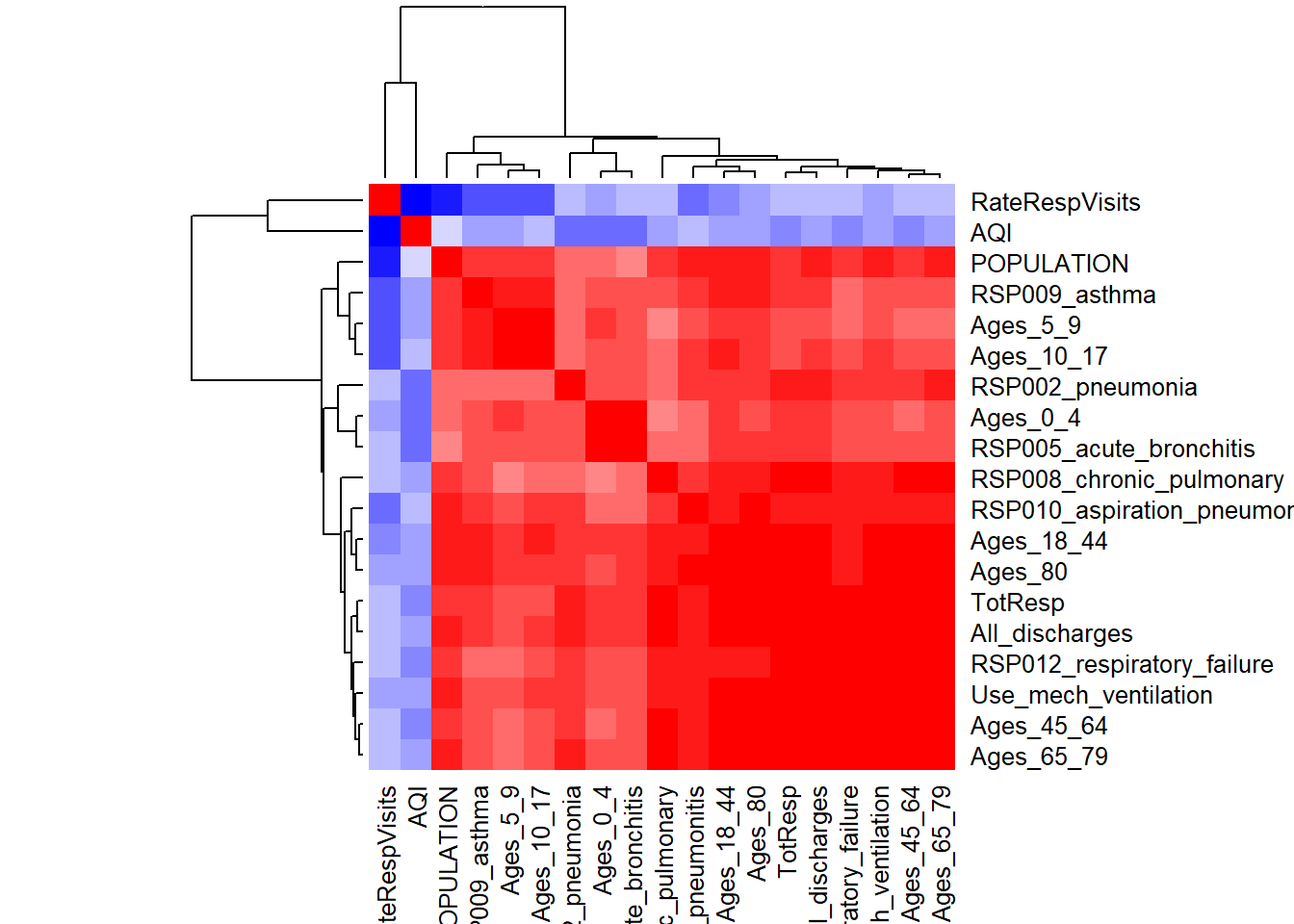
chart.Correlation(TotRespData[,9:26], histogram = T, pch = 19)



#### coefficients whose magnitude are between 0.3 and 0.5 indicate variables which have a low correlation

palette = colorRampPalette(c("blue", "white", "red")) (20)

heatmap(x = correlationMatrix, col = palette, symm = TRUE)

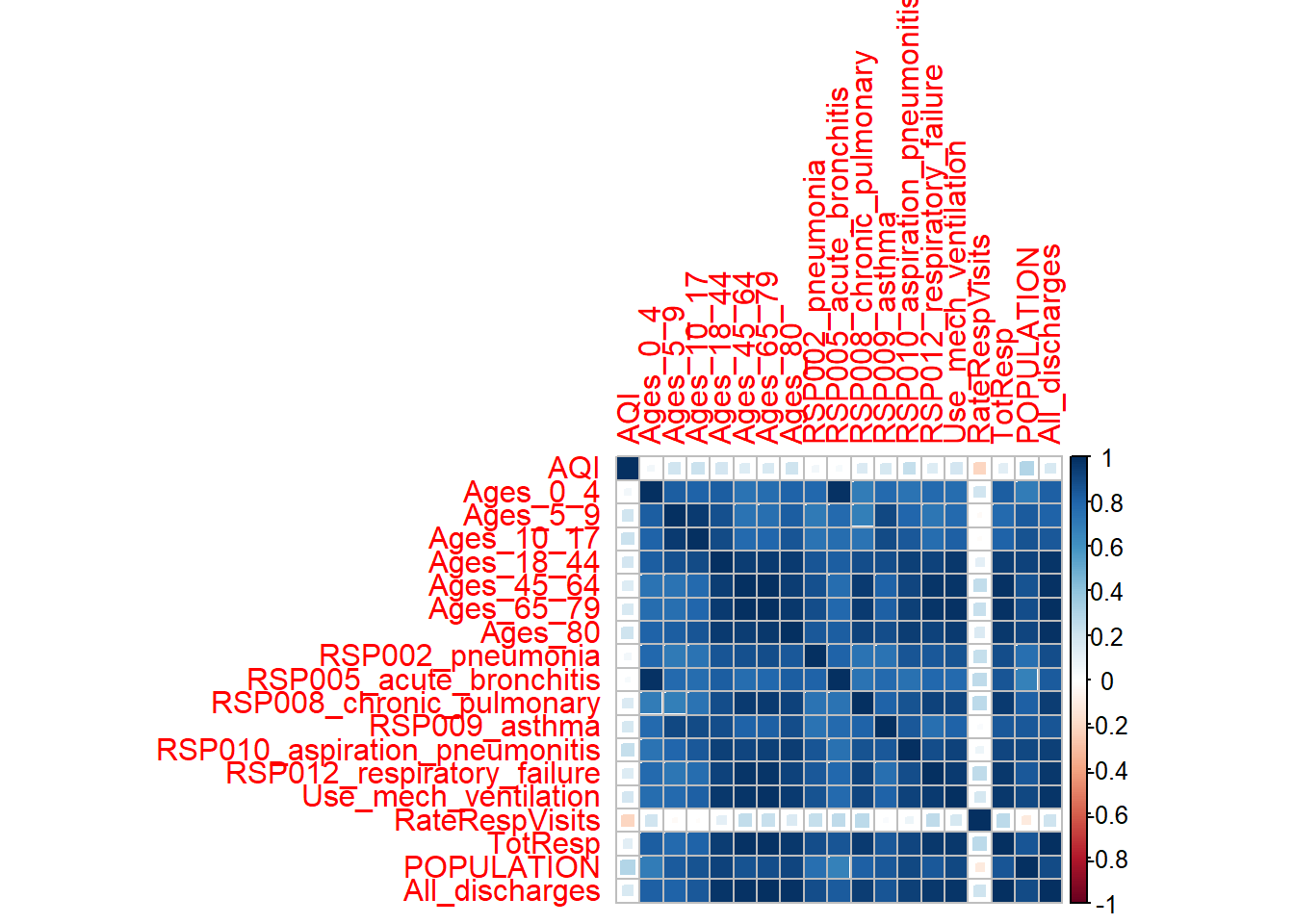
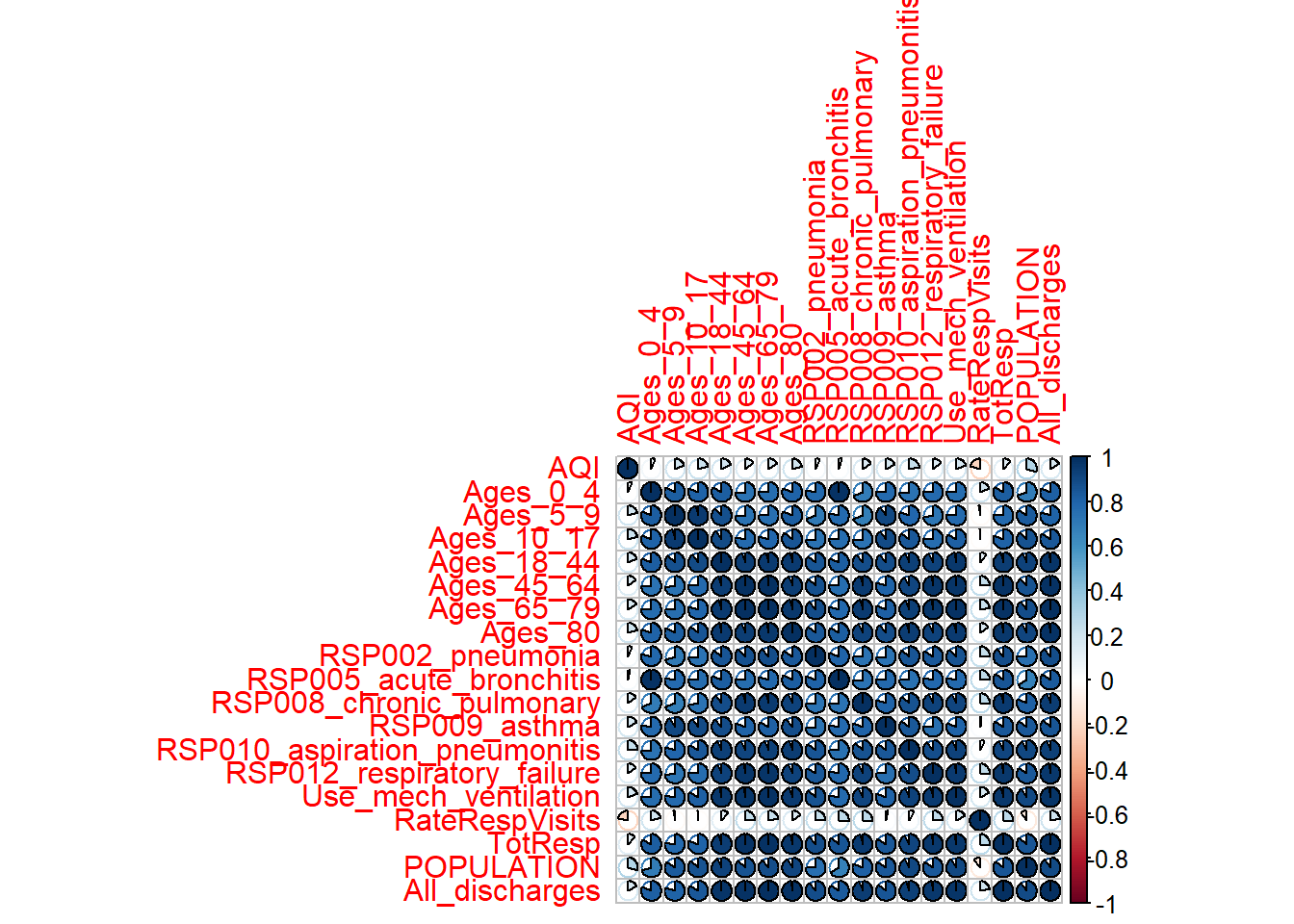
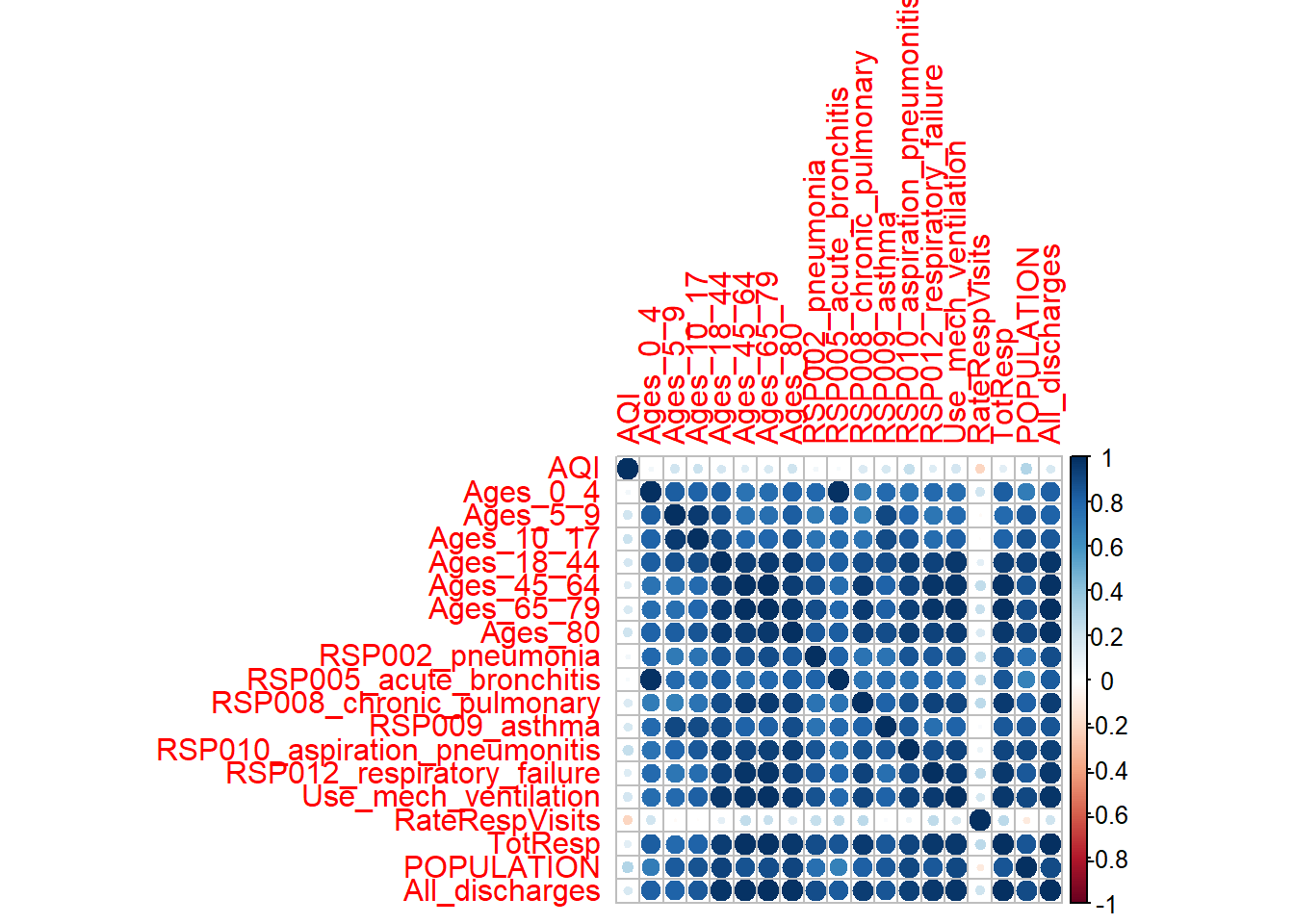


c1<-corrplot(correlationMatrix, method="circle")

c2<-corrplot(correlationMatrix, method="pie")

c3<-corrplot(correlationMatrix, method="number")

c4<-corrplot(correlationMatrix, method="square")



# Saving in a data file the Tibble with Hospital data Visits only with Respiratory ilnesses and AQI data by USA States:

save(TotRespData, file = paste(datapath,file\_TotRespData\_save,sep=""))

# BaseLine

#### Marcos Mariscal

#### 23/5/2021

#### This RMarkdown gaves first aproach with a lineal regression model

Getting only numeric Features for Respiratory illnesses

NumDataResp <- TotRespData[,8:26]

### Normalization the Data

#### Z-score normalization, consists of subtracting the mean and divide by the standard deviation

We normalize the data to bring all the variables to the same range

means <- colMeans(NumDataResp)

sds <- apply(NumDataResp, 2, sd)

z\_NumDataResp <- scale(NumDataResp, center = means, scale = sds)

z\_NumDataResp <- as\_tibble(z\_NumDataResp)

#### Splitting data into test and train

The following code splits 70% of the data selected randomly into training set and the remaining 30% sample into test data set

dt = sort(sample(nrow(NumDataResp), nrow(NumDataResp)\*.7))

train<-z\_NumDataResp[dt,]

test<-z\_NumDataResp[-dt,]

### BASE LINE.Linear Regression

The fist approach for a good prediction model y to get results of a basic model such a Linear Regression

linreg1<- lm(AQI~., data = train)

summary(linreg1)

##

## Call:

## lm(formula = AQI ~ ., data = train)

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.2045 -0.5320 -0.1232 0.3629 6.4873

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -0.008849 0.026204 -0.338 0.735669

## Ages\_0\_4 -0.854660 0.397127 -2.152 0.031602 \*

## Ages\_5\_9 0.051565 0.130029 0.397 0.691763

## Ages\_10\_17 0.021103 0.104697 0.202 0.840294

## Ages\_18\_44 -0.885483 0.249667 -3.547 0.000406 \*\*\*

## Ages\_45\_64 -1.075800 0.859036 -1.252 0.210709

## Ages\_65\_79 -1.645760 1.100117 -1.496 0.134938

## Ages\_80 -0.244704 0.683625 -0.358 0.720448

## RSP002\_pneumonia -0.001856 0.331508 -0.006 0.995534

## RSP005\_acute\_bronchitis 0.680179 0.353275 1.925 0.054437 .

## RSP008\_chronic\_pulmonary 0.665936 0.752570 0.885 0.376410

## RSP009\_asthma -0.027277 0.120046 -0.227 0.820293

## RSP010\_aspiration\_pneumonitis 0.339340 0.101157 3.355 0.000821 \*\*\*

## RSP012\_respiratory\_failure 0.775666 0.374186 2.073 0.038405 \*

## Use\_mech\_ventilation 0.185389 0.189070 0.981 0.327035

## RateRespVisits -0.028574 0.039208 -0.729 0.466292

## TotResp -2.573548 1.766614 -1.457 0.145460

## POPULATION 0.612056 0.130290 4.698 2.96e-06 \*\*\*

## All\_discharges 4.205001 3.027602 1.389 0.165142

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.8817 on 1123 degrees of freedom

## Multiple R-squared: 0.2212, Adjusted R-squared: 0.2087

## F-statistic: 17.72 on 18 and 1123 DF, p-value: < 2.2e-16

#### We have to coefficients smaller than 0.05 in Features Ages\_18\_44 and RPS010, then theses are good features to predict Bad Air Quality conditions.

Regarding R squared value is not good 0.20, because it refers that this model has good predictions only for the 20% of the values. Some variables could be taken out to for this model.

linreg2<- lm(AQI~Ages\_0\_4 +Ages\_10\_17+Ages\_18\_44+Ages\_45\_64+RSP002\_pneumonia+RSP010\_aspiration\_pneumonitis+RSP012\_respiratory\_failure+Use\_mech\_ventilation+ RateRespVisits+POPULATION, data = train)

summary(linreg2)

##

## Call:

## lm(formula = AQI ~ Ages\_0\_4 + Ages\_10\_17 + Ages\_18\_44 + Ages\_45\_64 +

## RSP002\_pneumonia + RSP010\_aspiration\_pneumonitis + RSP012\_respiratory\_failure +

## Use\_mech\_ventilation + RateRespVisits + POPULATION, data = train)

##

## Residuals:

## Min 1Q Median 3Q Max

## -3.2119 -0.5207 -0.1201 0.3626 6.4820

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -0.003438 0.026281 -0.131 0.895934

## Ages\_0\_4 -0.087150 0.060986 -1.429 0.153274

## Ages\_10\_17 0.017039 0.075788 0.225 0.822156

## Ages\_18\_44 -0.491977 0.165562 -2.972 0.003026 \*\*

## Ages\_45\_64 -0.648132 0.190660 -3.399 0.000699 \*\*\*

## RSP002\_pneumonia -0.305124 0.066179 -4.611 4.47e-06 \*\*\*

## RSP010\_aspiration\_pneumonitis 0.415483 0.088379 4.701 2.91e-06 \*\*\*

## RSP012\_respiratory\_failure 0.270971 0.127117 2.132 0.033250 \*

## Use\_mech\_ventilation 0.378167 0.161274 2.345 0.019206 \*

## RateRespVisits -0.012049 0.037508 -0.321 0.748087

## POPULATION 0.668349 0.105671 6.325 3.64e-10 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.8865 on 1131 degrees of freedom

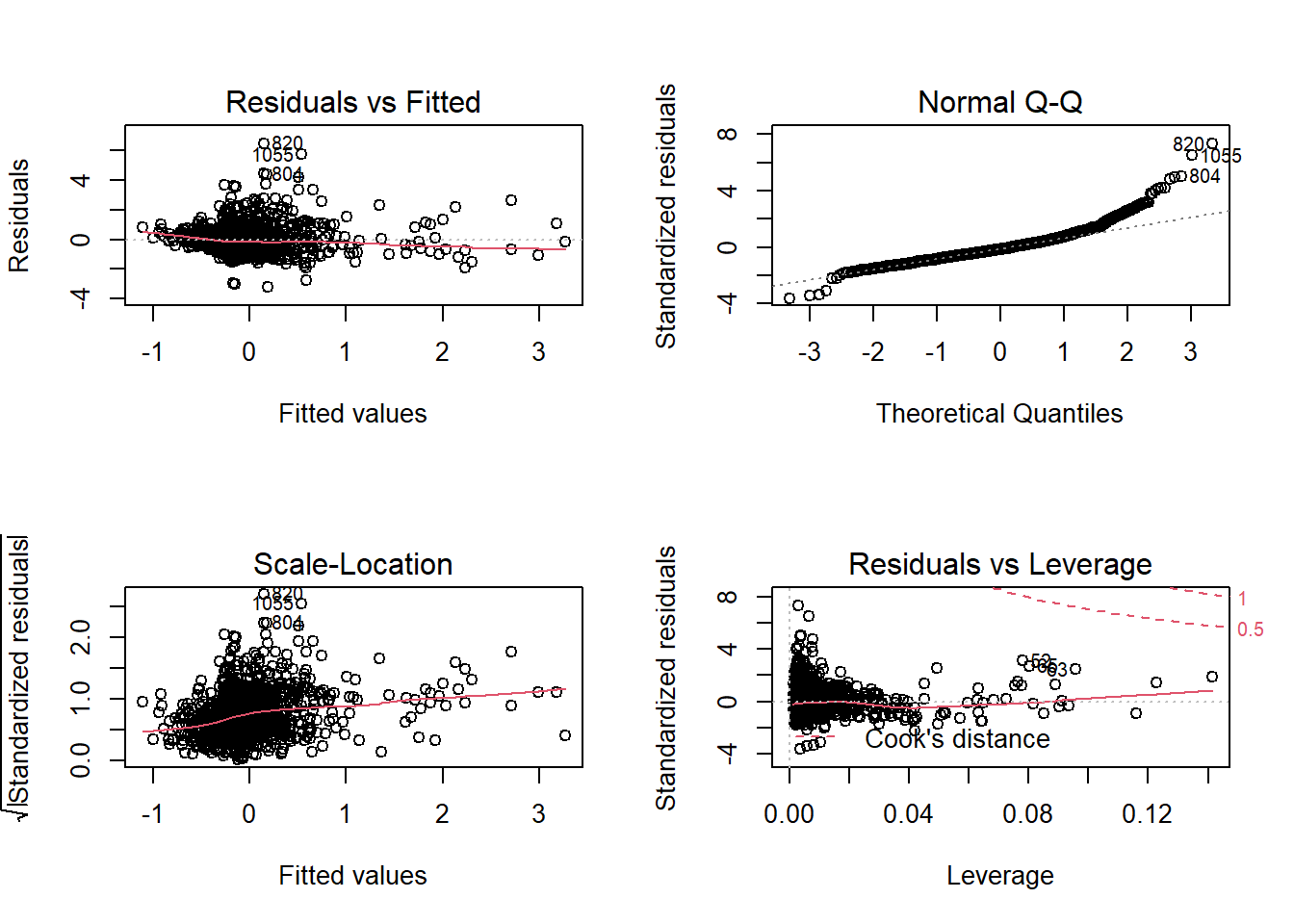
## Multiple R-squared: 0.2071, Adjusted R-squared: 0.2001

## F-statistic: 29.54 on 10 and 1131 DF, p-value: < 2.2e-16

### In this second model we can see than most of variables pass the Hypothesis Test

layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))

plot(linreg2)

It seems than linear regressor model is not good to fit our dataset

### Now we calculate Root Mean Square Error (RMSE) to assess how well

Predic1<- predict(linreg2, newdata= test)

rmse(test$AQI,Predic1)

## [1] 0.9349882

### RMSE value ≥0.5 reflects the poor ability of the model to accurately predict the data.

# Random Forest

#### Marcos Mariscal

#### 23/5/2021

#### This RMarkdown gaves second aproach with a Random Forest regression model

Getting only numeric Features for Respiratory ilnesses

NumDataResp <- TotRespData[,8:26]

#### Normalization the Data

Z-score normalization, consists of subtracting the mean and divide by the standard deviation We normalize the data to bring all the variables to the same range

means <- colMeans(NumDataResp)

sds <- apply(NumDataResp, 2, sd)

z\_NumDataResp <- scale(NumDataResp, center = means, scale = sds)

z\_NumDataResp <- as\_tibble(z\_NumDataResp)

####Splitting data into test and train

The following code splits 70% of the data selected randomly into training set and the remaining 30% sample into test data set

set.seed(111)

dt = sort(sample(nrow(NumDataResp), nrow(NumDataResp)\*.7))

train<-z\_NumDataResp[dt,]

test<-z\_NumDataResp[-dt,]

#+RSP008\_chronic\_pulmonary+RSP009\_asthma+RSP010\_aspiration\_pneumonitis+RSP012\_respiratory\_failure+Use\_mech\_ventilation+RateRespVisits+TotResp+POPULATION

set.seed(222)

rf <- randomForest(AQI ~ Ages\_0\_4+Ages\_5\_9+Ages\_10\_17+Ages\_18\_44+Ages\_45\_64+Ages\_65\_79+Ages\_80+RSP002\_pneumonia+RSP005\_acute\_bronchitis+RSP008\_chronic\_pulmonary+RSP009\_asthma+RSP010\_aspiration\_pneumonitis+RSP012\_respiratory\_failure+Use\_mech\_ventilation+RateRespVisits+TotResp, data = train,

mtry =10,

ntree = 50,

proximity=F)

summary(rf)

## Length Class Mode

## call 6 -none- call

## type 1 -none- character

## predicted 1142 -none- numeric

## mse 50 -none- numeric

## rsq 50 -none- numeric

## oob.times 1142 -none- numeric

## importance 16 -none- numeric

## importanceSD 0 -none- NULL

## localImportance 0 -none- NULL

## proximity 0 -none- NULL

## ntree 1 -none- numeric

## mtry 1 -none- numeric

## forest 11 -none- list

## coefs 0 -none- NULL

## y 1142 -none- numeric

## test 0 -none- NULL

## inbag 0 -none- NULL

## terms 3 terms call

print(rf)

##

## Call:

## randomForest(formula = AQI ~ Ages\_0\_4 + Ages\_5\_9 + Ages\_10\_17 + Ages\_18\_44 + Ages\_45\_64 + Ages\_65\_79 + Ages\_80 + RSP002\_pneumonia + RSP005\_acute\_bronchitis + RSP008\_chronic\_pulmonary + RSP009\_asthma + RSP010\_aspiration\_pneumonitis + RSP012\_respiratory\_failure + Use\_mech\_ventilation + RateRespVisits + TotResp, data = train, mtry = 10, ntree = 50, proximity = F)

## Type of random forest: regression

## Number of trees: 50

## No. of variables tried at each split: 10

##

## Mean of squared residuals: 0.8748744

## % Var explained: 18.22

predict1 <- predict(rf, test)

RMSE(predict1, test$AQI)

## [1] 0.8019525

### RMSE value ≥0.5 reflects the poor ability of the model to accurately predict the data

This RMSE value (0.91) is worse than the Base Line one (0.84)

## Now visualize original test and predicted data in a plot

pre <- predict1

act <- test$AQI

t1 <- cbind(pre,act)

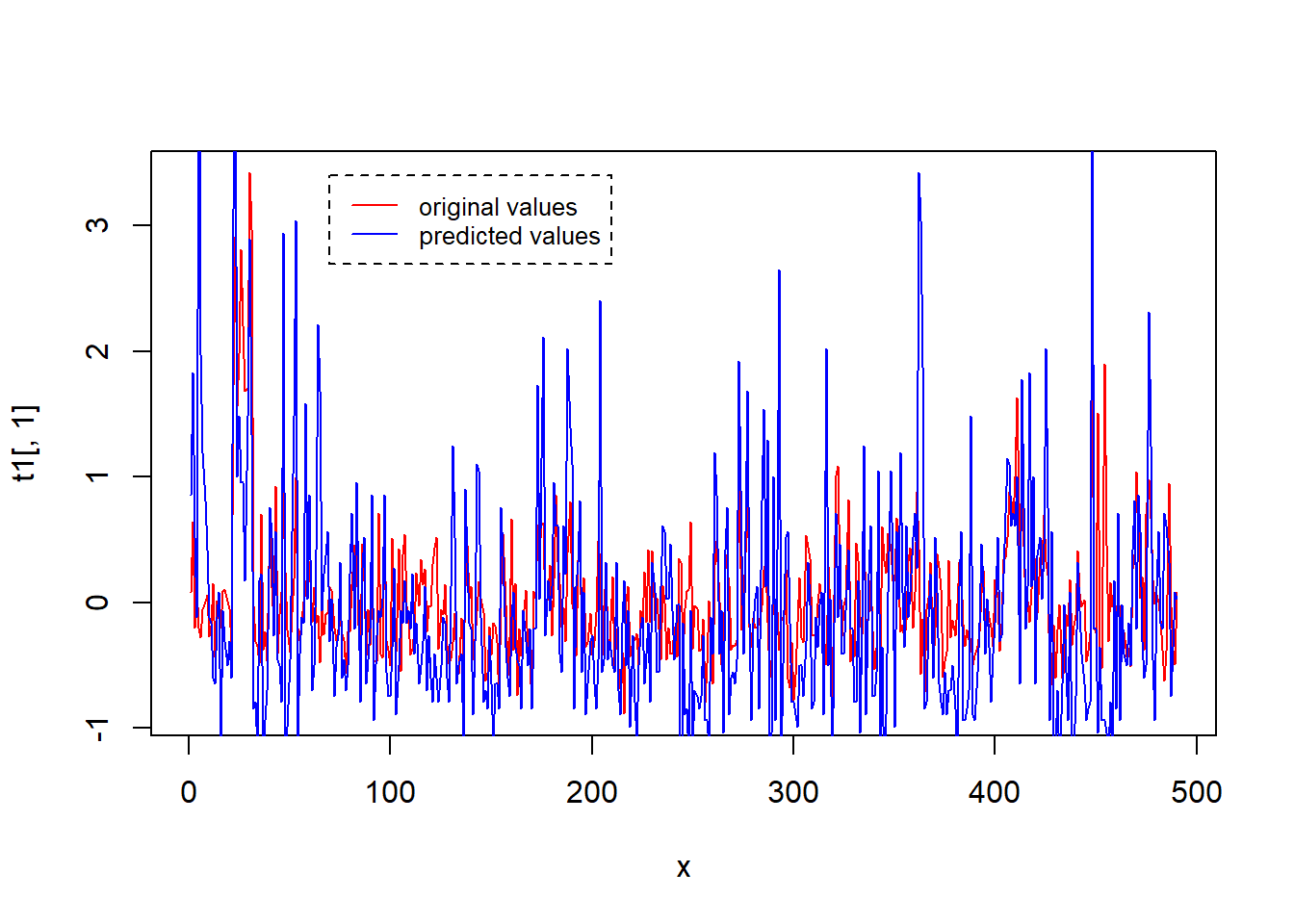
x = 1:length(t1[,1])

plot(x, t1[,1], col = "red", type = "l")

lines(x, t1[,2], col = "blue", type = "l")

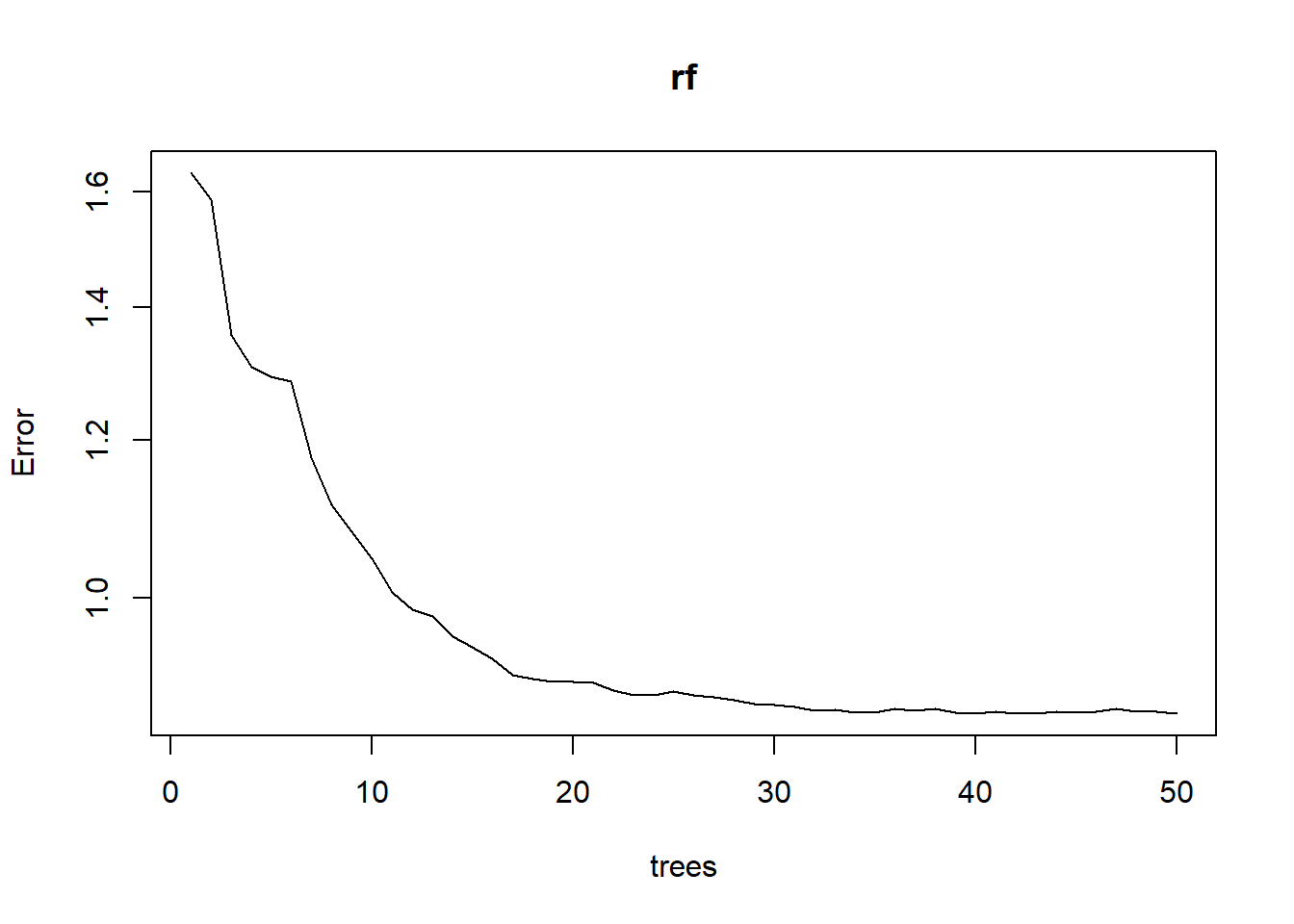
legend(x = 70, y=3.4, legend = c("original values", "predicted values"),

col = c("red", "blue"), box.lty = 2, cex = 0.8, lty = c(1, 1))



### Plotting Model

plot(rf, log="y")



# XGBoost

#### Marcos Mariscal

#### 23/5/2021

#### This RMarkdown gaves last aproach with a XGBoost regression model

Getting only numeric Features for Respiratory illnesses

# taking out features not related to respiratory illnesses

NumDataResp <- TotRespData[,8:24]

#### Normalization the Data

Z-score normalization, consists of subtracting the mean and divide by the standard deviation We normalize the data to bring all the variables to the same range

means <- colMeans(NumDataResp)

sds <- apply(NumDataResp, 2, sd)

z\_NumDataResp <- scale(NumDataResp, center = means, scale = sds)

z\_NumDataResp <- base::data.frame(z\_NumDataResp)

### Column AQI is excluded because it will be our label column, the one we want to predict.

df = z\_NumDataResp[,!(names(z\_NumDataResp) %in% "AQI")]

AQI\_vector = z\_NumDataResp$AQI

head(df)

## Ages\_0\_4 Ages\_5\_9 Ages\_10\_17 Ages\_18\_44 Ages\_45\_64 Ages\_65\_79

## 1 0.98178126 0.2694311 -0.12772069 0.35635600 0.0871224102 0.26855147

## 2 1.15378160 0.9341699 0.17802212 0.07750154 0.0007153399 0.24055033

## 3 0.36213899 0.2162519 -0.05128498 0.04325625 -0.0143119767 0.31129004

## 4 -0.08770807 0.2694311 -0.47168135 -0.08394052 -0.1720988008 0.12117707

## 5 -0.21560576 -0.2357705 -0.01306713 -0.08394052 -0.1580106915 0.01948873

## 6 -0.47140115 -0.4750765 0.13980427 -0.32365751 -0.3834204403 -0.21262595

## Ages\_80 RSP002\_pneumonia RSP005\_acute\_bronchitis RSP008\_chronic\_pulmonary

## 1 0.32213733 -0.2341323 1.1362252 0.32281397

## 2 0.20904592 -0.3948527 1.2062807 0.24543226

## 3 0.31747376 -0.4254661 0.3656147 0.24653771

## 4 0.04698711 -0.6014932 -0.1364498 -0.01324376

## 5 -0.05794306 -0.6627201 -0.2882367 -0.06741096

## 6 -0.21184063 -0.7392536 -0.5217550 -0.24759981

## RSP009\_asthma RSP010\_aspiration\_pneumonitis RSP012\_respiratory\_failure

## 1 0.46969001 -0.111412312 0.5282091

## 2 0.46969001 -0.064410918 0.6182206

## 3 -0.01829573 0.358601630 0.5424215

## 4 -0.08946031 0.170596053 0.4642536

## 5 -0.13012579 0.288099538 0.2605435

## 6 -0.41478414 0.006091173 -0.1705640

## Use\_mech\_ventilation RateRespVisits TotResp

## 1 0.20214172 0.6565932 0.42483032

## 2 0.29388123 0.5865740 0.39123541

## 3 0.34824538 0.2202416 0.21547071

## 4 0.12059548 -0.3114985 -0.03965589

## 5 0.09681117 -0.5246005 -0.14190128

## 6 -0.19539616 -1.0269123 -0.38290828

### Splitting data into test and train

The following code splits 70% of the data selected randomly into training set and the remaining 30% sample into test data set. train: will be used to build the model test: will be used to assess the quality of our model.

set.seed(100)

ind <- sample(2, nrow(df), replace = T, prob = c(.7, .3))

df\_train <- df[ind==1,1:16]

df\_test <- df[ind==2, 1:16]

t\_train <- setDT(df\_train)

t\_test <- setDT(df\_test)

labels <- df[ind==1, 16]

ts\_labels <- df[ind==2, 16]

dtrain <- xgb.DMatrix(label = labels, data = as.matrix(df\_train))

dtest <- xgb.DMatrix(label = ts\_labels, data = as.matrix(df\_test))

### Extreme Gradient Boosting (xgboost).

With linear model solver and tree learning algorithms.

eta=0.1; used in update to prevents overfitting (defult=0.3) max\_depth=15; the trees won’t be deep nthread = 2; the number of CPU threads we are going to use eval\_metric=“rmse”; Evaluation metrics for validation data, rmse for regression nrounds=150; the number of passes on the data. After these number of passes there is no rmse reduce

### Predicting

# predict

print(xgbmod)

## ##### xgb.Booster

## raw: 1.4 Kb

## call:

## xgb.train(params = params, data = dtrain, nrounds = nrounds,

## watchlist = watchlist, verbose = verbose, print\_every\_n = print\_every\_n,

## early\_stopping\_rounds = early\_stopping\_rounds, maximize = maximize,

## save\_period = save\_period, save\_name = save\_name, xgb\_model = xgb\_model,

## callbacks = callbacks, booster = "gblinear", eta = 0.1, max\_depth = 15,

## nthread = 2, eval\_metric = "rmse")

## params (as set within xgb.train):

## booster = "gblinear", eta = "0.1", max\_depth = "15", nthread = "2", eval\_metric = "rmse", validate\_parameters = "TRUE"

## xgb.attributes:

## niter

## callbacks:

## cb.print.evaluation(period = print\_every\_n)

## cb.evaluation.log()

## # of features: 16

## niter: 500

## nfeatures : 16

## evaluation\_log:

## iter train\_rmse

## 1 0.537659

## 2 0.446335

## ---

## 499 0.030293

## 500 0.030280

predic <- predict(xgbmod, dtest)

print(length(predic))

## [1] 518

RMSE(predic, ts\_labels)

## [1] 0.02573248

#### Great RMSE value, less than 0.03

### Now visualize original test and predicted data in a plot

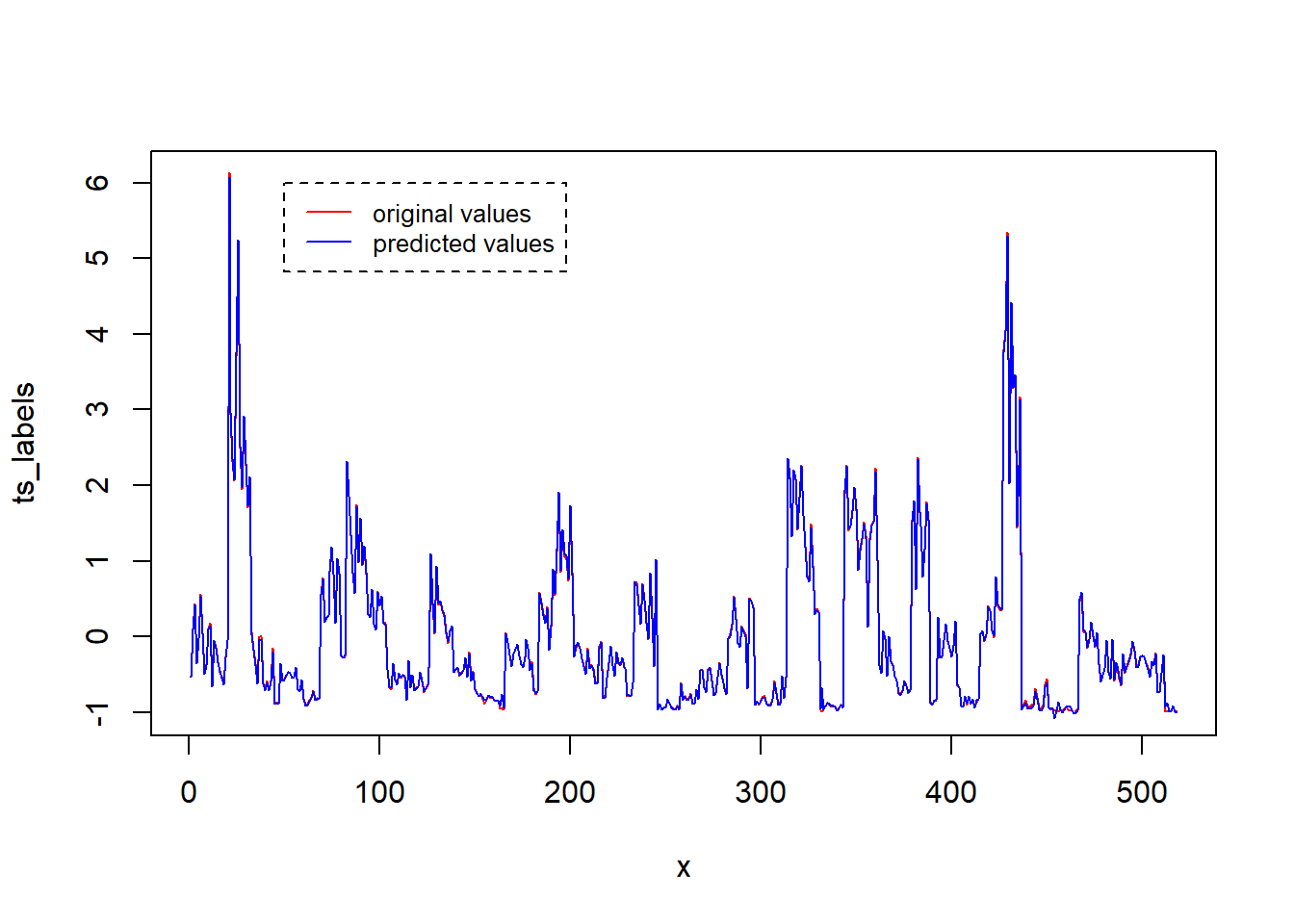
x = 1:length(ts\_labels)

plot(x, ts\_labels, col = "red", type = "l")

lines(x, predic, col = "blue", type = "l")

legend(x = 50, y=6, legend = c("original values", "predicted values"),

col = c("red", "blue"), box.lty = 2, cex = 0.8, lty = c(1, 1))



### Now we can see the Features importance

# Feature Importance

importance\_matrix <- xgb.importance(model = xgbmod)

print(importance\_matrix)

## Feature Weight

## 1: RSP008\_chronic\_pulmonary 0.24028200

## 2: TotResp 0.22711400

## 3: RSP012\_respiratory\_failure 0.13927500

## 4: Ages\_65\_79 0.11022400

## 5: RSP002\_pneumonia 0.10825900

## 6: Ages\_45\_64 0.10556000

## 7: RSP005\_acute\_bronchitis 0.09354080

## 8: Ages\_80 0.07757350

## 9: Use\_mech\_ventilation -0.06403430

## 10: RSP009\_asthma 0.03945460

## 11: Ages\_18\_44 -0.03161280

## 12: Ages\_0\_4 0.02028470

## 13: RSP010\_aspiration\_pneumonitis -0.01758310

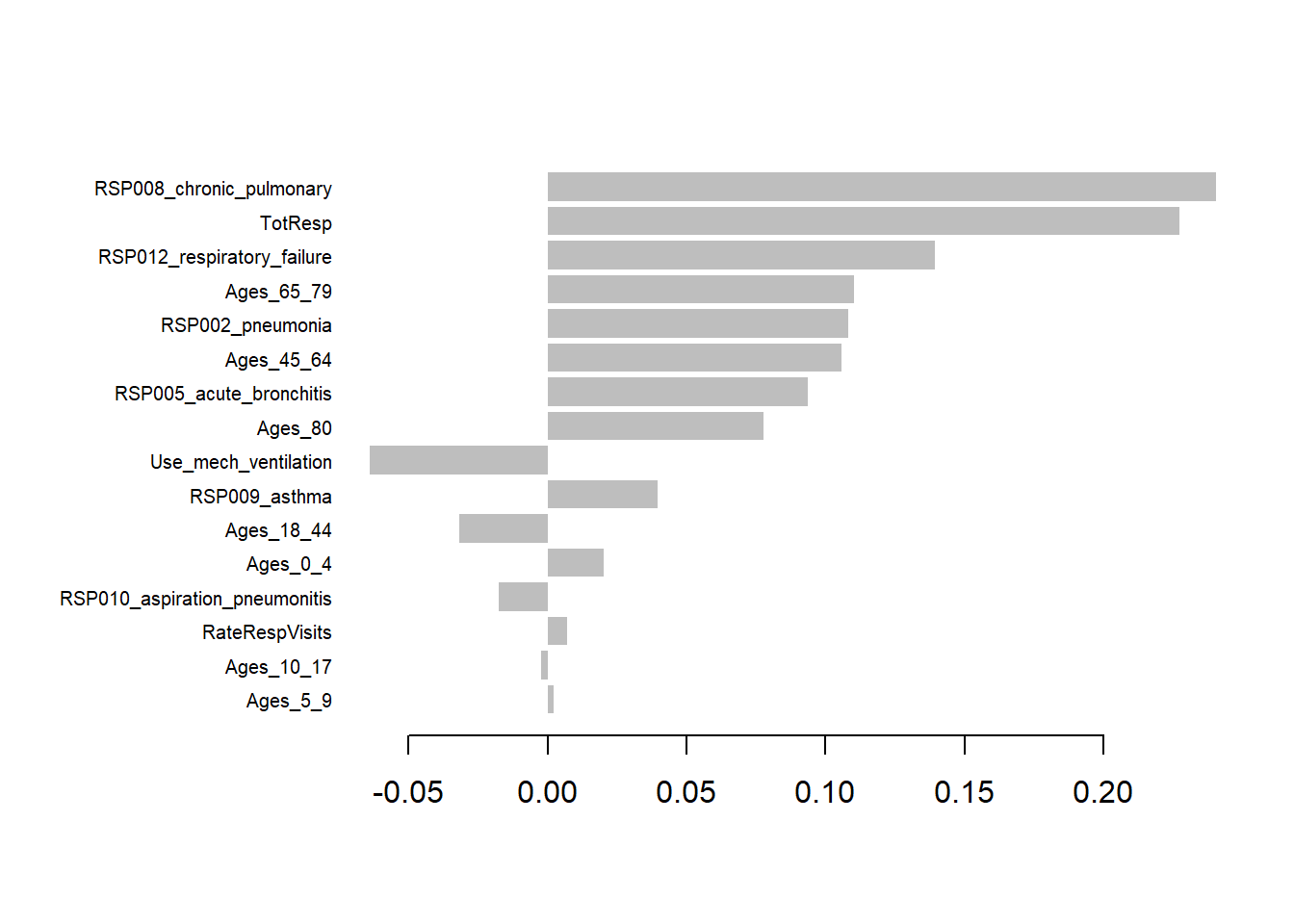
## 14: RateRespVisits 0.00694655

## 15: Ages\_10\_17 -0.00215977

## 16: Ages\_5\_9 0.00211791

### Plot Importance Matrix

xgb.plot.importance(importance\_matrix = importance\_matrix)



#save model

saveRDS(xgbmod, file = "xgbmod.rds")

xgbmod<-readRDS("xgbmod.rds")